Justin Solomon

Numerical Algorithms

In memory of Clifford Nass
(1958-2013)

Contents

SECTION I **Preliminaries**

CHAPTER 0 ▪ Mathematics Review 3

0.1 PRELIMINARIES: NUMBERS AND SETS 3
0.2 VECTOR SPACES 4
0.2.1 Defining Vector Spaces 4
0.2.2 Span, Linear Independence, and Bases 5
0.2.3 Our Focus: $\mathbb{R}^n$ 7
0.3 LINEARITY 9
0.3.1 Matrices 10
0.3.2 Scalars, Vectors, and Matrices 12
0.3.3 Matrix Storage and Multiplication Methods 13
0.3.4 Model Problem: $A\vec{x} = \vec{b}$ 14
0.4 NON-LINEARITY: DIFFERENTIAL CALCULUS 15
0.4.1 Differentiation in One Variable 15
0.4.2 Differentiation in Multiple Variables 17
0.4.3 Optimization 20
0.5 EXERCISES 22

CHAPTER 1 ▪ Numerics and Error Analysis 25

1.1 STORING NUMBERS WITH FRACTIONAL PARTS 25
1.1.1 Fixed Point Representations 26
1.1.2 Floating Point Representations 27
1.1.3 More Exotic Options 29
1.2 UNDERSTANDING ERROR 30
1.2.1 Classifying Error 31
1.2.2 Conditioning, Stability, and Accuracy 32
1.3 PRACTICAL ASPECTS 34
1.3.1 Computing Vector Norms 34
1.3.2 Larger-Scale Example: Summation 35
1.4 EXERCISES 36

SECTION II Linear Algebra

CHAPTER 2 ▪ Linear Systems and the LU Decomposition 43

2.1 SOLVABILITY OF LINEAR SYSTEMS 43
2.2 AD-HOC SOLUTION STRATEGIES 45
2.3 ENCODING ROW OPERATIONS 47
2.3.1 Permutation 47
2.3.2 Row Scaling 48
2.3.3 Elimination 48
2.4 GAUSSIAN ELIMINATION 51
2.4.1 Forward Substitution 51
2.4.2 Back Substitution 52
2.4.3 Analysis of Gaussian Elimination 53
2.5 LU FACTORIZATION 55
2.5.1 Constructing the Factorization 56
2.5.2 Using the Factorization 56
2.5.3 Implementing LU 57
2.6 EXERCISES 58

CHAPTER 3 ▪ Designing and Analyzing Linear Systems 61

3.1 SOLUTION OF SQUARE SYSTEMS 61
3.1.1 Regression 62
3.1.2 Least Squares 64
3.1.3 Tikhonov Regularization 65
3.1.4 Image Alignment 66
3.1.5 Deconvolution 68
3.1.6 Harmonic Parameterization 69
3.2 SPECIAL PROPERTIES OF LINEAR SYSTEMS 70
3.2.1 Positive Definite Matrices and the Cholesky Factorization 70
3.2.2 Sparsity 74
3.2.3 Additional Special Structures 75
3.3 SENSITIVITY ANALYSIS 76
3.3.1 Matrix and Vector Norms 76
3.3.2 Condition Numbers 79
3.4 EXERCISES 81

CHAPTER 4 ▪ Column Spaces and QR 87

4.1 THE STRUCTURE OF THE NORMAL EQUATIONS 87
4.2 ORTHOGONALITY 88

4.3 STRATEGY FOR NON-ORTHOGONAL MATRICES 89
4.4 GRAM-SCHMIDT ORTHOGONALIZATION 90
4.4.1 Projections 90
4.4.2 Gram-Schmidt Algorithm 92
4.5 HOUSEHOLDER TRANSFORMATIONS 94
4.6 REDUCED QR FACTORIZATION 96
4.7 EXERCISES 97

CHAPTER 5 ▪ Eigenvectors 101

5.1 MOTIVATION 101
5.1.1 Statistics 102
5.1.2 Differential Equations 103
5.1.3 Spectral Embedding 104
5.2 PROPERTIES OF EIGENVECTORS 106
5.2.1 Symmetric and Positive Definite Matrices 108
5.2.2 Specialized Properties 110
5.2.2.1 Characteristic Polynomial 110
5.2.2.2 Jordan Normal Form 110
5.3 COMPUTING A SINGLE EIGENVALUE 111
5.3.1 Power Iteration 111
5.3.2 Inverse Iteration 112
5.3.3 Shifting 113
5.4 FINDING MULTIPLE EIGENVALUES 114
5.4.1 Deflation 114
5.4.2 QR Iteration 115
5.4.3 Krylov Subspace Methods 119
5.5 SENSITIVITY AND CONDITIONING 120
5.6 EXERCISES 121

CHAPTER 6 ▪ Singular Value Decomposition 125

6.1 DERIVING THE SVD 125
6.1.1 Computing the SVD 127
6.2 APPLICATIONS OF THE SVD 128
6.2.1 Solving Linear Systems and the Pseudoinverse 128
6.2.2 Decomposition into Outer Products and Low-Rank Approximations 129
6.2.3 Matrix Norms 130
6.2.4 The Procrustes Problem and Point Cloud Alignment 131
6.2.5 Principal Component Analysis (PCA) 133

6.2.6 Eigenfaces 133
6.3 EXERCISES 135

SECTION III **Nonlinear Techniques**

CHAPTER 7 ▪ Nonlinear Systems 141

7.1 ROOT-FINDING IN A SINGLE VARIABLE 141
7.1.1 Characterizing Problems 141
7.1.2 Continuity and Bisection 142
7.1.3 Fixed Point Iteration 143
7.1.4 Newton's Method 145
7.1.5 Secant Method 146
7.1.6 Hybrid Techniques 147
7.1.7 Single-Variable Case: Summary 148
7.2 MULTIVARIABLE PROBLEMS 148
7.2.1 Newton's Method 149
7.2.2 Making Newton Faster: Quasi-Newton and Broyden 149
7.3 CONDITIONING 152
7.4 EXERCISES 152

CHAPTER 8 ▪ Unconstrained Optimization 155

8.1 UNCONSTRAINED OPTIMIZATION: MOTIVATION 155
8.2 OPTIMALITY 157
8.2.1 Differential Optimality 158
8.2.2 Alternative Conditions for Optimality 160
8.3 ONE-DIMENSIONAL STRATEGIES 161
8.3.1 Newton's Method 162
8.3.2 Golden Section Search 162
8.4 MULTIVARIABLE STRATEGIES 165
8.4.1 Gradient Descent 166
8.4.2 Newton's Method in Multiple Variables 166
8.4.3 Optimization without Hessians: BFGS 167
8.5 EXERCISES 169
8.6 APPENDIX: DERIVATION OF BFGS UPDATE 174

CHAPTER 9 ▪ Constrained Optimization 177

9.1 MOTIVATION 178
9.2 THEORY OF CONSTRAINED OPTIMIZATION 181
9.2.1 Optimality 181

9.2.2 KKT Conditions 181
9.3 OPTIMIZATION ALGORITHMS 184
9.3.1 Sequential Quadratic Programming (SQP) 185
9.3.1.1 Equality constraints 185
9.3.1.2 Inequality Constraints 185
9.3.2 Barrier Methods 186
9.4 CONVEX PROGRAMMING 186
9.4.1 Linear Programming 188
9.4.2 Second-Order Cone Programming 189
9.4.3 Semidefinite Programming 191
9.4.4 Integer Programs and Relaxations 193
9.5 EXERCISES 194

CHAPTER 10 ▪ Iterative Linear Solvers 199

10.1 GRADIENT DESCENT 200
10.1.1 Gradient Descent for Linear Systems 200
10.1.2 Convergence 201
10.2 CONJUGATE GRADIENTS 203
10.2.1 Motivation 204
10.2.2 Suboptimality of Gradient Descent 205
10.2.3 Generating A-Conjugate Directions 207
10.2.4 Formulating the Conjugate Gradients Algorithm 208
10.2.5 Convergence and Stopping Conditions 211
10.3 PRECONDITIONING 211
10.3.1 CG with Preconditioning 212
10.3.2 Common Preconditioners 213
10.4 OTHER ITERATIVE SCHEMES 214
10.5 EXERCISES 215

CHAPTER 11 ▪ Specialized Optimization Methods 219

11.1 NONLINEAR LEAST SQUARES 219
11.1.1 Gauss-Newton 220
11.1.2 Levenberg-Marquardt 220
11.2 ITERATIVELY-REWEIGHTED LEAST SQUARES 221
11.3 COORDINATE DESCENT AND ALTERNATION 223
11.3.1 Identifying Candidates for Alternation 223
11.3.2 Augmented Lagrangians and ADMM 227
11.4 GLOBAL OPTIMIZATION 232
11.4.1 Graduated Optimization 233

11.4.2 Randomized Global Optimization 235
11.5 ONLINE OPTIMIZATION 237
11.6 EXERCISES 240

SECTION IV **Functions, Derivatives, and Integrals**

CHAPTER 12 ▪ Interpolation 249

12.1 INTERPOLATION IN A SINGLE VARIABLE 250
12.1.1 Polynomial Interpolation 250
12.1.2 Alternative Bases 253
12.1.3 Piecewise Interpolation 255
12.2 MULTIVARIABLE INTERPOLATION 257
12.2.1 Nearest-Neighbor Interpolation 257
12.2.2 Barycentric Interpolation 258
12.2.3 Grid-Based Interpolation 260
12.3 THEORY OF INTERPOLATION 260
12.3.1 Linear Algebra of Functions 261
12.3.2 Approximation via Piecewise Polynomials 263
12.4 EXERCISES 264

CHAPTER 13 ▪ Integration and Differentiation 269

13.1 MOTIVATION 270
13.2 QUADRATURE 271
13.2.1 Interpolatory Quadrature 272
13.2.2 Quadrature Rules 273
13.2.3 Newton-Cotes Quadrature 274
13.2.4 Gaussian Quadrature 278
13.2.5 Adaptive Quadrature 279
13.2.6 Multiple Variables 281
13.2.7 Conditioning 282
13.3 DIFFERENTIATION 282
13.3.1 Differentiating Basis Functions 283
13.3.2 Finite Differences 283
13.3.3 Richardson Extrapolation 285
13.3.4 Choosing the Step Size 286
13.3.5 Automatic Differentiation 287
13.3.6 Integrated Quantities and Structure Preservation 288
13.4 EXERCISES 289

CHAPTER 14 ▪ Ordinary Differential Equations 293

14.1 MOTIVATION 294
14.2 THEORY OF ODES 295
14.2.1 Basic Notions 295
14.2.2 Existence and Uniqueness 299
14.2.3 Model Equations 299
14.3 TIME-STEPPING SCHEMES 301
14.3.1 Forward Euler 301
14.3.2 Backward Euler 303
14.3.3 Trapezoidal Method 304
14.3.4 Runge-Kutta Methods 305
14.3.5 Exponential Integrators 307
14.4 MULTIVALUE METHODS 308
14.4.1 Newmark Schemes 308
14.4.2 Staggered Grid and Leapfrog 311
14.5 COMPARISON OF INTEGRATORS 312
14.6 EXERCISES 314

CHAPTER 15 ▪ Partial Differential Equations 319

15.1 MOTIVATION 320
15.2 STATEMENT AND STRUCTURE OF PDES 325
15.2.1 Properties of PDEs 325
15.2.2 Boundary Conditions 326
15.3 MODEL EQUATIONS 328
15.3.1 Elliptic PDEs 328
15.3.2 Parabolic PDEs 329
15.3.3 Hyperbolic PDEs 330
15.4 REPRESENTING DERIVATIVE OPERATORS 331
15.4.1 Finite Differences 332
15.4.2 Collocation 336
15.4.3 Finite Elements 337
15.4.4 Finite Volumes 340
15.4.5 Other Methods 341
15.5 SOLVING PARABOLIC AND HYPERBOLIC EQUATIONS 342
15.5.1 Semidiscrete Methods 342
15.5.2 Fully Discrete Methods 343
15.6 NUMERICAL CONSIDERATIONS 344
15.6.1 Consistency, Convergence, and Stability 344
15.6.2 Linear Solvers for PDE 345
15.7 EXERCISES 345

Preface

COMPUTER science (CS) is experiencing a fundamental shift in its approach to modeling and problem-solving. Early computer scientists primarily studied *discrete* mathematics. With the introduction of fast floating-point processing alongside "big data," three-dimensional scanning, and other sources of noisy input, however, practitioners of CS now must design robust methods for processing and understanding real-valued data.

Numerical Algorithms introduces the skills necessary to be both *clients* and *designers* of numerical methods for computer science applications. This text is designed for students who are comfortable with mathematical notation and formality but need to review continuous concepts alongside the algorithms under consideration. It covers a broad base of topics, from numerical linear algebra to optimization and differential equations, with the goal of deriving standard approaches while developing sufficient intuition and comfort to approach more extensive literature in each subtopic. Thus, in each chapter, we will gently but rigorously introduce numerical methods alongside mathematical background and motivating examples from modern problems in computer science.

Most chapters begin and end with real-world motivation for why someone might use a given class of numerical algorithms. For example, the singular value decomposition is introduced alongside statistical methods, point cloud alignment, and low-rank approximation, and discussion of least-squares includes concepts from machine learning like kernelization and regularization. The goal of this presentation of theory and application in parallel is to improve intuition for the design of numerical methods and the application of each method to practical problems.

Special care has been taken to provide unifying threads from chapter to chapter. This strategy helps unify discussions of seemingly independent problems, reinforcing skills while presenting increasingly complex algorithms. In particular, starting with a chapter on mathematical preliminaries, methods are introduced with *variational* principles in mind, e.g. solving the linear system $A\vec{x} = \vec{b}$ by minimizing the energy $\|A\vec{x} - \vec{b}\|_2^2$ or finding eigenvectors as critical points of the Rayleigh quotient.

Numerical algorithms are very different from algorithms approached in most other branches of computer science, and students should expect to be challenged the first time they study this material. With practice, however, it can be easy to build up intuition for this unique and widely-applicable field.

The refinement of course notes and other materials leading to this textbook benefited from generous input from my students and colleagues. In the interests of maintaining these materials and responding to the needs of students and instructors, please do not hesitate to contact me with questions, comments, concerns, or ideas for potential changes.

Justin Solomon

Acknowledgments

PREPARATION of this textbook would not have been possible without the support of countless individuals and organizations. I have attempted to acknowledge some of the many contributors and supporters below. I cannot thank these colleagues and friends enough for their patience and attention throughout this undertaking.

The book is dedicated to the memory of Prof. Clifford Nass, whose guidance fundamentally shaped my early academic career. His wisdom, teaching, encouragement, enthusiasm, and unique sense of style all will be missed on the Stanford campus and in the larger community.

My mother, Nancy Griesemer, was the first to suggest expanding my teaching materials into a text. I also would not have been able to find the time or energy to prepare this work without her support or that from my father Rodney Solomon, my sister Julia Solomon Ensor and her husband Jeff Ensor, and grandmothers Juddy Solomon and Dolores Griesemer. My uncle Peter Silberman and aunt Dena Silberman have supported my academic career from its inception. Many other family members also should be thanked including Archa and Joseph Emerson; Jerry, Jinny, Kate, Bonnie, and Jeremiah Griesemer; Jim, Marge, Paul, Laura, Jarrett, Liza, Jiana, Lana, Jahson, Jaime, Gabriel, and Jesse Solomon; Chuck and Louise Silverberg; and Barbara, Kerry, Greg, and Amy Schaner.

My career at Stanford has been guided primarily by my advisor Leonidas Guibas and co-advisor Adrian Butscher. The approaches I take to many of the problems in the book undoubtedly imitate the problem-solving strategies they have taught me. Ron Fedkiw suggested I teach the course leading to this text and provided advice on preparing the material. My collaborators in the Geometric Computing Group and elsewhere on campus—including Mirela Ben-Chen, Tanya Glozman, Jonathan Huang, Qixing Huang, Michael Kerber, Andy Nguyen, Maks Ovsjanikov, Franco Pestilli, Chris Piech, Raif Rustamov, and Fan Wang — kindly have allowed me to use some research time to complete this text and have helped refine the discussion at many points. Staff in the Stanford computer science department, including Meredith Hutchin, Claire Stager, and Steven Magness, made it possible to organize my numerical algorithms course and many others.

I owe many thanks to the students of Stanford's CS 205A course (fall 2013) for catching numerous typos and mistakes in an early draft of this book. The following is a no-doubt incomplete list of students and course assistants who contributed to this effort: Scott Chung, Tao Du, Lennart Jansson, Miles Johnson, David Hyde, Luke Knepper, Minjae Lee, Nisha Masharani, David McLaren, Catherine Mullings, John Reyna, William Song, Ben-Han Sung, Martina Troesch, Ozhan Turgut, Patrick Ward, Joongyeub Yeo, and Yang Zhao.

David Hyde and Scott Chung continued to provide detailed feedback in winter and spring 2014. In addition, they helped prepare figures and end-of-chapter problems. Problems that they drafted are marked DH and SC, resp.

I leaned upon several colleagues and friends to help edit the text. In addition to those mentioned above, additional contributors include: Nick Alger, George Anderson, Rahil Baber, Nicolas Bonneel, Chen Chen, Matthew Cong, Roy Frostig, Jessica Hwang, Howon Lee, Julian Kates-Harbeck, Jonathan Lee, Niru Maheswaranathan, Mark Pauly, and Dan Robinson.

Special thanks to Jan Heiland and Tao Du for helping clarify the derivation of the BFGS algorithm.

Sarah Chow, Randi Cohen, and Kate Gallo at Taylor & Francis guided me through the publication process and answered countless questions as I prepared this work for print.

The Hertz Foundation provided a valuable network of experienced and knowledgeable members of the academic community. In particular, Louis Lerman provided career advice throughout my PhD that shaped my approach to research and navigating academia. Other members of the Hertz community who provided guidance include Diann Callaghan, Wendy Cieslak, Jay Davis, Philip Eckhoff, Linda Kubiak, Amanda O'Connor, Linda Souza, Thomas Weaver, and Katherine Young. I should also acknowledge the NSF GRFP and NDSEG fellowships for their support.

A multitude of friends supported this work in assorted stages of its development. Additional collaborators and mentors in the research community who have discussed and encouraged this work include Keenan Crane, Michael Eichmair, Hao Li, Niloy Mitra, Helmut Pottmann, Fei Sha, Olga Sorkine-Hornung, Amir Vaxman, Etienne Vouga, Brian Wandell, and Chris Wojtan. The first several chapters of this book were drafted on tour with the Stanford Symphony Orchestra on their European tour "In Beethoven's Footsteps" (summer 2013). Beyond this tour, Geri Actor, Susan Bratman, Debra Fong, Stephen Harrison, Patrick Kim, Mindy Perkins, Thomas Shoebotham, and Lowry Yankwich all supported musical breaks during the drafting of this book. Prometheus Athletics provided an unexpected outlet, and I should thank Archie de Torres, Amy Giver, Lori Giver, Troy Obrero, and Ben Priestley for allowing me to be an enthusiastic if clumsy participant.

Additional friends who have lent advice, assistance, and time to this effort include: Chris Aakre, Katy Ashe, Katya Avagian, Kyle Barrett, Noelle Beegle, Gilbert Bernstein, Elizabeth Blaber, Lia Bonamassa, Eric Boromisa, Karen Budd, Arun Chaganty, Phil Chen, Andrew Chou, Bernie Chu, Cindy Chu, Victor Cruz, Elan Dagenais, Abe Davis, Matthew Decker, Martin Duncan, Eric Ellenoff, James Estrella, Alyson Falwell, Anna French, Adair Gerke, Christina Goeders, Gabrielle Gulo, Nathan Hall-Snyder, Logan Hehn, Jo Jaffe, Dustin Janatpour, Brandon Johnson, Victoria Johnson, Jeff Gilbert, Stephanie Go, Alex Godofsky, Alan Guo, Randy Hernando, Petr Johanes, Maria Judnick, Ken Kao, Jonathan Kass, Gavin Kho, Sarah Kongpachith, Jim Lalonde, Lauren Lax, Atticus Lee, Eric Lee, Menyoung Lee, Letitia Lew, Siyang Li, Adrian Lim, Yongwhan Lim, Alex Louie, Lily Louie, Cleo Messinger, Courtney Meyer, Daniel Meyer, Lisa Newman, Logan Obrero, Pualani Obrero, Thomas Obrero, Molly Pam, David Parker, Madeline Paymer, Cuauhtemoc Peranda, Fabianna Perez, Bharath Ramsundar, Arty Rivera, Daniel Rosenfeld, Te Rutherford, Ravi Sankar, Aaron Sarnoff, Amanda Schloss, Keith Schwarz, Steve Sellers, Charlton Soesanto, Mark Smitt, Jacob Steinhardt, Charlie Syms, Michael Tamkin, Sumil Thapa, Herb Tyson, Katie Tyson, Madeleine Udell, Greg Valdespino, Walter Vulej, Thomas Waggoner, Frank Wang, Sydney Wang, Susanna Wen, Genevieve Williams, Molby Wong, Winston Yan, and Evan Young.

I

Preliminaries

CHAPTER 0

Mathematics Review

CONTENTS

0.1 Preliminaries: Numbers and Sets 3
0.2 Vector Spaces 4
0.2.1 Defining Vector Spaces 4
0.2.2 Span, Linear Independence, and Bases 5
0.2.3 Our Focus: $\mathbb{R}^n$ 7
0.3 Linearity 9
0.3.1 Matrices 10
0.3.2 Scalars, Vectors, and Matrices 12
0.3.3 Matrix Storage and Multiplication Methods 13
0.3.4 Model Problem: $A\vec{x} = \vec{b}$ 14
0.4 Non-Linearity: Differential Calculus 15
0.4.1 Differentiation in One Variable 15
0.4.2 Differentiation in Multiple Variables 17
0.4.3 Optimization 20

IN this chapter we will review notions from linear algebra and multivariable calculus that will be relevant to our discussion of computational techniques. It is intended as a review of background material with a bias toward ideas and interpretations commonly encountered in practice; the chapter safely can be skipped or used as reference by students with stronger background in mathematics.

0.1 PRELIMINARIES: NUMBERS AND SETS

Rather than considering algebraic (and at times philosophical) discussions like "What is a number?," we will rely on intuition and mathematical common sense to define a few sets:

- The *natural numbers* $\mathbb{N} = \{1, 2, 3, \ldots\}$
- The *integers* $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
- The *rational numbers* $\mathbb{Q} = \{a/b : a, b \in \mathbb{Z}, b \neq 0\}$
- The *real numbers* $\mathbb{R}$ encompassing $\mathbb{Q}$ as well as *irrational* numbers like π and $\sqrt{2}$
- The *complex numbers* $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$, where we think of i as satisfying $i = \sqrt{-1}$.

The definition of $\mathbb{Q}$ is the first of many times that we will use the notation $\{A : B\}$; the braces denote a set and the colon can be read as "such that." For instance, the definition of $\mathbb{Q}$ can be read as "the set of fractions a/b *such that* a and b are integers;" as a second

example, we could write $\mathbb{N} = \{n \in \mathbb{Z} : n > 0\}$. It is worth acknowledging that our definition of $\mathbb{R}$ is far from rigorous. The construction of the real numbers can be an important topic for practitioners of cryptography techniques that make use of alternative number systems, but these intricacies are irrelevant for the discussion at hand.

As with any other sets, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ can be manipulated using generic operations to generate new sets of numbers. In particular, we can define the "Euclidean product" of two sets A and B as

$$A \times B = \{(a, b) : a \in A \text{ and } b \in B\}.$$

We can take *powers* of sets by writing

$$A^n = \underbrace{A \times A \times \cdots \times A}_{n \text{ times}}.$$

This construction yields what will become our favorite set of numbers in chapters to come:

$$\mathbb{R}^n = \{(a_1, a_2, \ldots, a_n) : a_i \in \mathbb{R} \text{ for all } i\}$$

0.2 VECTOR SPACES

Introductory linear algebra courses easily could be titled "Introduction to Finite-Dimensional Vector Spaces." Although the definition of a vector space might appear abstract, we will find many concrete applications that all satisfy the formal aspects and thus can benefit from the machinery we will develop.

0.2.1 Defining Vector Spaces

We begin by defining a vector space and providing a number of examples:

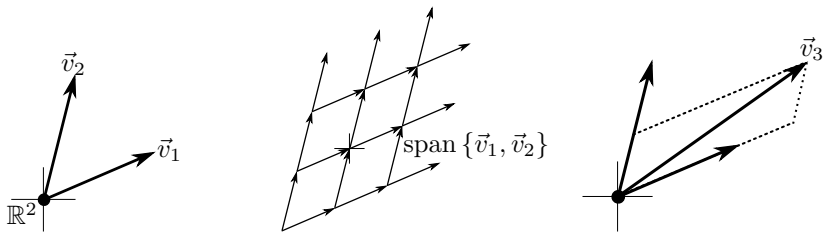
Definition 0.1 (Vector space over $\mathbb{R}$)**.** A *vector space* over $\mathbb{R}$ is a set $\mathcal{V}$ closed under addition and scalar multiplication satisfying the following axioms:

- Additive commutativity and associativity: For all $\vec{u}, \vec{v}, \vec{w} \in \mathcal{V}$, $\vec{v} + \vec{w} = \vec{w} + \vec{v}$ and $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$.
- Distributivity: For all $\vec{v}, \vec{w} \in \mathcal{V}$ and $a, b \in \mathbb{R}$, $a(\vec{v} + \vec{w}) = a\vec{v} + a\vec{w}$ and $(a + b)\vec{v} = a\vec{v} + b\vec{v}$.
- Additive identity: There exists $\vec{0} \in \mathcal{V}$ with $\vec{0} + \vec{v} = \vec{v}$ for all $\vec{v} \in \mathcal{V}$.
- Additive inverse: For all $\vec{v} \in \mathcal{V}$, there exists $\vec{w} \in \mathcal{V}$ with $\vec{v} + \vec{w} = \vec{0}$.
- Multiplicative identity: For all $\vec{v} \in \mathcal{V}$, $1 \cdot \vec{v} = \vec{v}$.
- Multiplicative compatibility: For all $\vec{v} \in \mathcal{V}$ and $a, b \in \mathbb{R}$, $(ab)\vec{v} = a(b\vec{v})$.

For our purposes, a scalar is a number in $\mathbb{R}$; a *complex* vector space satisfies the same definition with $\mathbb{R}$ replaced by $\mathbb{C}$. It is usually straightforward to spot vector spaces in the wild, including the following examples:

Example 0.1 ($\mathbb{R}^n$ as a vector space)**.** The most common example of a vector space is $\mathbb{R}^n$. Here, addition and scalar multiplication happen component-by-component:

$$(1, 2) + (-3, 4) = (1 - 3, 2 + 4) = (-2, 6)$$
$$10 \cdot (-1, 1) = (10 \cdot -1, 10 \cdot 1) = (-10, 10)$$



(a) $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$ (b) span $\{\vec{v}_1, \vec{v}_2\}$ (c) span $\{\vec{v}_1, \vec{v}_2, \vec{v}_3\}$

FIGURE 0.1 (a) Two vectors $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^2$; (b) their span is the whole plane $\mathbb{R}^2$; (c) span $\{\vec{v}_1, \vec{v}_2, \vec{v}_3\}$ = span $\{\vec{v}_1, \vec{v}_2\}$ because $\vec{v}_3$ can be written as a linear combination of $\vec{v}_1$ and $\vec{v}_2$.

Example 0.2 (Polynomials)**.** A second example of a vector space is the ring of polynomials with real-valued coefficients, denoted $\mathbb{R}[x]$. A polynomial $p \in \mathbb{R}[x]$ is a function $p : \mathbb{R} \to \mathbb{R}$ taking the form*

$$p(x) = \sum_k a_k x^k.$$

Addition and scalar multiplication are carried out in the usual way, e.g. if $p(x) = x^2 + 2x - 1$ and $q(x) = x^3$, then $3p(x) + 5q(x) = 5x^3 + 3x^2 + 6x - 3$, which is another polynomial. As an aside, for future examples note that functions like $p(x) = (x-1)(x+1) + x^2(x^3-5)$ are still polynomials even though they are not explicitly written in the form above.

Elements $\vec{v} \in \mathcal{V}$ of a vector space $\mathcal{V}$ are called *vectors*, and a weighted sum of the form $\sum_i a_i \vec{v}_i$, where $a_i \in \mathbb{R}$ and $\vec{v}_i \in \mathcal{V}$, is known as a *linear combination* of the $\vec{v}_i$'s. In our second example, the "vectors" are polynomials, although we do not normally use this language to discuss $\mathbb{R}[x]$. One way to link these two viewpoints would be to identify the polynomial $\sum_k a_k x^k$ with the sequence $(a_0, a_1, a_2, \ldots)$; remember that polynomials have finite numbers of terms, so the sequence eventually will end in a string of zeros.

0.2.2 Span, Linear Independence, and Bases

Suppose we start with vectors $\vec{v}_1, \ldots, \vec{v}_k \in \mathcal{V}$ for vector space $\mathcal{V}$. By Definition 0.1, we have two ways to start with these vectors and construct new elements of $\mathcal{V}$: addition and scalar multiplication. The idea of *span* is that it describes all of the vectors you can reach via these two operations:

Definition 0.2 (Span)**.** The *span* of a set $S \subseteq \mathcal{V}$ of vectors is the set

$$\text{span}\, S \equiv \{a_1\vec{v}_1 + \cdots + a_k\vec{v}_k : \vec{v}_i \in \mathcal{V} \text{ and } a_i \in \mathbb{R} \text{ for all } i\}.$$

Figure 0.1(b) illustrates the span of two vectors shown in Figure 0.1(a). Notice that span S is a *subspace* of $\mathcal{V}$, that is, a subset of $\mathcal{V}$ that is in itself a vector space. We can provide a few examples:

*The notation $f : A \to B$ means f is a function that takes as input an element of set A and outputs an element of set B. For instance, $f : \mathbb{R} \to \mathbb{Z}$ takes as input a real number in $\mathbb{R}$ and outputs an integer $\mathbb{Z}$, as might be the case for $f(x) = \lfloor x \rfloor$, the "round down" function.

Example 0.3 (Mixology)**.** The typical well at a cocktail bar contains at least four ingredients at the bartender's disposal: vodka, tequila, orange juice, and grenadine. Assuming we have this well, we can represent drinks as points in $\mathbb{R}^4$, with one element for each ingredient. For instance, a typical tequila sunrise can be represented using the point $(0, 1.5, 6, 0.75)$, representing amounts of vodka, tequila, orange juice, and grenadine (in ounces), resp.

The set of drinks that can be made with the typical well is contained in

$$\text{span}\{(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1)\},$$

that is, all combinations of the four basic ingredients. A bartender looking to save time, however, might notice that many drinks have the same orange juice to grenadine ratio and mix the bottles. The new simplified well may be easier for pouring but can make fundamentally fewer drinks:

$$\text{span}\{(1,0,0,0),(0,1,0,0),(0,0,6,0.75)\}$$

For example, this reduced well cannot fulfill orders for a screwdriver, which contains orange juice but not grenadine.

Example 0.4 (Polynomials)**.** Define $p_k(x) \equiv x^k$. Then, we can write

$$\mathbb{R}[x] = \text{span}\{p_k : k \geq 0\}.$$

Make sure you understand notation well enough to see why this is the case.

Adding another item to a set of vectors does not always increase the size of its span, as illustrated in Figure 0.1(c). For instance, in $\mathbb{R}^2$,

$$\text{span}\{(1,0),(0,1)\} = \text{span}\{(1,0),(0,1),(1,1)\}.$$

In this case, we say that the set $\{(1,0),(0,1),(1,1)\}$ is *linearly dependent*:

Definition 0.3 (Linear dependence)**.** We provide three equivalent definitions. A set $S \subseteq \mathcal{V}$ of vectors is *linearly dependent* if:

1. One of the elements of S can be written as a linear combination of the other elements, or S contains zero.
2. There exists a non-empty linear combination of elements $\vec{v}_k \in \mathcal{S}$ yielding $\sum_{k=1}^{m} c_k \vec{v}_k = 0$ where $c_k \neq \vec{0}$ for all k.
3. There exists $\vec{v} \in S$ such that $\text{span}\, S = \text{span}\, S \backslash \{\vec{v}\}$. That is, we can remove a vector from S without affecting its span.

If S is not linearly dependent, then we say it is *linearly independent*.

Providing proof or informal evidence that each definition is equivalent to its counterparts (in an "if and only if" fashion) is a worthwhile exercise for students less comfortable with notation and abstract mathematics.

The concept of linear dependence leads to an idea of "redundancy" in a set of vectors. In this sense, it is natural to ask how large a set we can choose before adding another vector cannot possibly increase the span. More specifically, suppose we have a linearly independent set $S \subseteq \mathcal{V}$, and now we choose an additional vector $\vec{v} \in \mathcal{V}$. Adding $\vec{v}$ to S leads to one of two possible outcomes:

1. The span of $S \cup \{\vec{v}\}$ is *larger* than the span of S.
2. Adding $\vec{v}$ to S has no effect on the span.

The *dimension* of $\mathcal{V}$ counts the number of times we can get the first outcome while building up a set of vectors.

Definition 0.4 (Dimension and basis)**.** The *dimension* of $\mathcal{V}$ is the maximal size $|S|$ of a linearly-independent set $S \subset \mathcal{V}$ such that $\operatorname{span} S = \mathcal{V}$. Any set S satisfying this property is called a *basis* for $\mathcal{V}$.

Example 0.5 ($\mathbb{R}^n$)**.** The *standard basis* for $\mathbb{R}^n$ is the set of vectors of the form

$$\vec{e}_k \equiv (\underbrace{0, \ldots, 0,}_{k-1 \text{ elements}} \; 1, \; \underbrace{0, \ldots, 0}_{n-k \text{ elements}}).$$

That is, $\vec{e}_k$ has all zeros except for a single one in the k-th element. It is clear that these vectors are linearly independent and form a basis; for example in $\mathbb{R}^3$ any vector (a, b, c) can be written as $a\vec{e}_1 + b\vec{e}_2 + c\vec{e}_3$. Thus, the dimension of $\mathbb{R}^n$ is n, as we would expect.

Example 0.6 (Polynomials)**.** It is clear that the set $\{1, x, x^2, x^3, \ldots\}$ is a linearly independent set of polynomials spanning $\mathbb{R}[x]$. This set is infinitely large, and thus the dimension of $\mathbb{R}[x]$ is ∞.

0.2.3 Our Focus: $\mathbb{R}^n$

Of particular importance for our purposes is the vector space $\mathbb{R}^n$, the so-called *n-dimensional Euclidean space*. This is nothing more than the set of coordinate axes encountered in high school math classes:

- $\mathbb{R}^1 \equiv \mathbb{R}$ is the number line
- $\mathbb{R}^2$ is the two-dimensional plane with coordinates (x, y)
- $\mathbb{R}^3$ represents three-dimensional space with coordinates (x, y, z)

Nearly all methods in this course will deal with transformations and functions on $\mathbb{R}^n$.

For convenience, we usually write vectors in $\mathbb{R}^n$ in "column form," as follows

$$(a_1, \ldots, a_n) \equiv \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

This notation will include vectors as special cases of *matrices* discussed below.

Unlike some vector spaces, $\mathbb{R}^n$ has not only a vector space structure, but also one additional construction that makes all the difference: the *dot product*.

Definition 0.5 (Dot product)**.** The dot product of two vectors $\vec{a} = (a_1, \ldots, a_n)$ and $\vec{b} = (b_1, \ldots, b_n)$ in $\mathbb{R}^n$ is given by

$$\vec{a} \cdot \vec{b} = \sum_{k=1}^{n} a_k b_k.$$

Example 0.7 ($\mathbb{R}^2$)**.** The dot product of $(1, 2)$ and $(-2, 6)$ is $1 \cdot -2 + 2 \cdot 6 = -2 + 12 = 10$.

The dot product is an example of a *metric*, and its existence gives a notion of geometry to $\mathbb{R}^n$. For instance, we can use the Pythagorean theorem to define the *norm* or *length* of a vector $\vec{a}$ as the square root

$$\|\vec{a}\|_2 \equiv \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\vec{a} \cdot \vec{a}}.$$

Then, the distance between two points $\vec{a}, \vec{b} \in \mathbb{R}^n$ is $\|\vec{b} - \vec{a}\|_2$.

Dot products yield not only notions of lengths and distances but also of angles. The following trigonometric identity holds for $\vec{a}, \vec{b} \in \mathbb{R}^3$:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\|_2 \|\vec{b}\|_2 \cos\theta$$

where θ is the angle between $\vec{a}$ and $\vec{b}$. For $n \geq 4$, however, the notion of "angle" is much harder to visualize for $\mathbb{R}^n$. We might *define* the angle θ between $\vec{a}$ and $\vec{b}$ to be the value θ given by

$$\theta \equiv \arccos \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2 \|\vec{b}\|_2}.$$

We must do our homework before making such a definition! In particular, cosine outputs values in the interval $[-1, 1]$, so we must check that the input to arc cosine (also notated $\cos^{-1}$) is in this interval; thankfully, the well-known Cauchy-Schwarz inequality $\vec{a} \cdot \vec{b} \leq \|\vec{a}\|_2 \|\vec{b}\|_2$ guarantees exactly this property.

When $\vec{a} = c\vec{b}$ for some $c \in \mathbb{R}$, we have $\theta = \arccos 1 = 0$, as we would expect: The angle between parallel vectors is zero. What does it mean for (nonzero) vectors to be perpendicular? Let's substitute $\theta = 90°$. Then, we have

$$0 = \cos 90° = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2 \|\vec{b}\|_2}.$$

Multiplying both sides by $\|\vec{a}\|_2 \|\vec{b}\|_2$ motivates the definition:

Definition 0.6 (Orthogonality)**.** Two vectors are perpendicular, or *orthogonal*, when $\vec{a} \cdot \vec{b} = 0$.

This definition is somewhat surprising from a geometric standpoint. We have managed to define what it means to be perpendicular without any explicit use of angles.

Aside 0.1. There are many theoretical questions to ponder here, some of which we will address in future chapters when they are more motivated:

- Do all vector spaces admit dot products or similar structures?
- Do all finite-dimensional vector spaces admit dot products?
- What might be a reasonable dot product between elements of $\mathbb{R}[x]$?

Intrigued students can consult texts on real and functional analysis.

0.3 LINEARITY

A function from one vector space to another that preserves linear structure is known as a *linear* function:

Definition 0.7 (Linearity)**.** Suppose $\mathcal{V}$ and $\mathcal{V}'$ are vector spaces. Then, $\mathcal{L} : \mathcal{V} \to \mathcal{V}'$ is *linear* if it satisfies the following two criteria for all $\vec{v}, \vec{v}_1, \vec{v}_2 \in \mathcal{V}$ and $c \in \mathbb{R}$:

- $\mathcal{L}$ preserves sums: $\mathcal{L}[\vec{v}_1 + \vec{v}_2] = \mathcal{L}[\vec{v}_1] + \mathcal{L}[\vec{v}_2]$
- $\mathcal{L}$ preserves scalar products: $\mathcal{L}[c\vec{v}] = c\mathcal{L}[\vec{v}]$

It is easy to generate linear maps between vector spaces, as we can see in the following examples:

Example 0.8 (Linearity in $\mathbb{R}^n$)**.** The following map $f : \mathbb{R}^2 \to \mathbb{R}^3$ is linear:

$$f(x, y) = (3x, 2x + y, -y)$$

We can check linearity as follows:

- Sum preservation:

$$\begin{aligned} f(x_1 + x_2, y_1 + y_2) &= (3(x_1 + x_2), 2(x_1 + x_2) + (y_1 + y_2), -(y_1 + y_2)) \\ &= (3x_1, 2x_1 + y_1, -y_1) + (3x_2, 2x_2 + y_2, -y_2) \\ &= f(x_1, y_1) + f(x_2, y_2) \quad \checkmark \end{aligned}$$

- Scalar product preservation:

$$\begin{aligned} f(cx, cy) &= (3cx, 2cx + cy, -cy) \\ &= c(3x, 2x + y, -y) \\ &= cf(x, y) \quad \checkmark \end{aligned}$$

Contrastingly, $g(x, y) \equiv xy^2$ is not linear. For instance, $g(1, 1) = 1$ but $g(2, 2) = 8 \neq 2 \cdot g(1, 1)$, so this form does not preserve scalar products.

Example 0.9 (Integration)**.** The following "functional" $\mathcal{L}$ from $\mathbb{R}[x]$ to $\mathbb{R}$ is linear:

$$\mathcal{L}[p(x)] \equiv \int_0^1 p(x)\, dx.$$

This somewhat more abstract example maps polynomials $p(x)$ to real numbers $\mathcal{L}[p(x)]$. For example, we can write

$$\mathcal{L}[3x^2 + x - 1] = \int_0^1 (3x^2 + x - 1)\, dx = \frac{1}{2}.$$

Linearity comes from the following well-known facts from calculus:

$$\int_0^1 c \cdot f(x)\, dx = c \int_0^1 f(x)\, dx$$
$$\int_0^1 [f(x) + g(x)]\, dx = \int_0^1 f(x)\, dx + \int_0^1 g(x)\, dx$$

We can write a particularly nice form for linear maps on $\mathbb{R}^n$. The vector $\vec{a} = (a_1, \ldots, a_n)$ is equal to the sum $\sum_k a_k \vec{e}_k$, where $\vec{e}_k$ is the k-th standard basis vector from Example 0.5. Then, if $\mathcal{L}$ is linear we know:

$$\begin{aligned}\mathcal{L}[\vec{a}] &= \mathcal{L}\left[\sum_k a_k \vec{e}_k\right] \text{ for the standard basis } \vec{e}_k \\ &= \sum_k \mathcal{L}\left[a_k \vec{e}_k\right] \text{ by sum preservation} \\ &= \sum_k a_k \mathcal{L}\left[\vec{e}_k\right] \text{ by scalar product preservation}\end{aligned}$$

This derivation shows:

$\mathcal{L}$ is completely determined by its action on the standard basis vectors $\vec{e}_k$.

That is, for any vector $\vec{a} \in \mathbb{R}^n$, we can use the sum above to determine $\mathcal{L}[\vec{a}]$ by linearly combining $\mathcal{L}[\vec{e}_1], \ldots, \mathcal{L}[\vec{e}_n]$.

Example 0.10 (Expanding a linear map)**.** Recall the map in Example 0.8 given by $f(x, y) = (3x, 2x + y, -y)$. We have $f(\vec{e}_1) = f(1, 0) = (3, 2, 0)$ and $f(\vec{e}_2) = f(0, 1) = (0, 1, -1)$. Thus, the formula above shows:

$$f(x, y) = xf(\vec{e}_1) + yf(\vec{e}_2) = x \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

0.3.1 Matrices

The expansion of linear maps above suggests one of many contexts in which it is useful to store multiple vectors in the same structure. More generally, say we have n vectors $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^m$. We can write each as a column vector:

$$\vec{v}_1 = \begin{pmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{m1} \end{pmatrix}, \vec{v}_2 = \begin{pmatrix} v_{12} \\ v_{22} \\ \vdots \\ v_{m2} \end{pmatrix}, \cdots, \vec{v}_n = \begin{pmatrix} v_{1n} \\ v_{2n} \\ \vdots \\ v_{mn} \end{pmatrix}$$

Carrying these around separately can be cumbersome notationally, so to simplify matters we combine them into a single $m \times n$ matrix:

$$\begin{pmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{pmatrix}$$

We will call the space of such matrices $\mathbb{R}^{m \times n}$.

Example 0.11 (Identity matrix)**.** We can store the standard basis for $\mathbb{R}^n$ in the $n \times n$ "identity matrix" $I_{n\times n}$ given by:

$$I_{n\times n} \equiv \begin{pmatrix} | & | & & | \\ \vec{e}_1 & \vec{e}_2 & \cdots & \vec{e}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Since we constructed matrices as convenient ways to store sets of vectors, we can use multiplication to express how they can be combined linearly. In particular, a matrix in $\mathbb{R}^{m\times n}$ can be multiplied by a column vector in $\mathbb{R}^n$ as follows:

$$\begin{pmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \equiv c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n$$

Expanding this sum yields the following explicit formula for matrix-vector products:

$$\begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} c_1v_{11} + c_2v_{12} + \cdots + c_nv_{1n} \\ c_1v_{21} + c_2v_{22} + \cdots + c_nv_{2n} \\ \vdots \\ c_1v_{m1} + c_2v_{m2} + \cdots + c_nv_{mn} \end{pmatrix}$$

Example 0.12 (Identity matrix multiplication)**.** For any $\vec{x} \in \mathbb{R}^n$, we can write $\vec{x} = I_{n\times n}\vec{x}$, where $I_{n\times n}$ is the identity matrix from Example 0.11.

Example 0.13 (Linear map)**.** We return once again to the expression from Example 0.8 to show one more alternative form:

$$f(x, y) = \begin{pmatrix} 3 & 0 \\ 2 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

We similarly define a product between a matrix in $M \in \mathbb{R}^{m\times n}$ and another matrix in $\mathbb{R}^{n\times p}$ by concatenating individual matrix-vector products:

$$M \begin{pmatrix} | & | & & | \\ \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \\ | & | & & | \end{pmatrix} \equiv \begin{pmatrix} | & | & & | \\ M\vec{c}_1 & M\vec{c}_2 & \cdots & M\vec{c}_n \\ | & | & & | \end{pmatrix}$$

Example 0.14 (Mixology)**.** Continuing Example 0.3, suppose we make a tequila sunrise and second concoction with equal parts of the two liquors in our simplified well. To find out how much of the basic ingredients are contained in each order, we could combine the recipes for each column-wise and use matrix multiplication:

$$\begin{array}{r} \\ \text{Vodka} \\ \text{Tequila} \\ \text{OJ} \\ \text{Grenadine} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Well 1} & \text{Well 2} & \text{Well 3} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \\ 0 & 0 & 0.75 \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cc} \text{Drink 1} & \text{Drink 2} \end{array} \\ \begin{pmatrix} 0 & 0.75 \\ 1.5 & 0.75 \\ 1 & 2 \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cc} \text{Drink 1} & \text{Drink 2} \end{array} \\ \begin{pmatrix} 0 & 0.75 \\ 1.5 & 0.75 \\ 6 & 12 \\ 0.75 & 1.5 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{Vodka} \\ \text{Tequila} \\ \text{OJ} \\ \text{Grenadine} \end{array}$$

We will use capital letters to represent matrices, like $A \in \mathbb{R}^{m\times n}$. We will use the notation $A_{ij} \in \mathbb{R}$ to denote the element of A at row i and column j.

0.3.2 Scalars, Vectors, and Matrices

If we wish to unify our notation completely, we can write a scalar as a 1×1 vector $c \in \mathbb{R}^{1\times 1}$. Similarly, as we already suggested in §0.2.3, if we write vectors in $\mathbb{R}^n$ in column form, they can be considered $n \times 1$ matrices $\vec{v} \in \mathbb{R}^{n\times 1}$. Matrix-vector products can be interpreted easily in this context; for example, if $A \in \mathbb{R}^{m\times n}$, $\vec{x} \in \mathbb{R}^n$, and $\vec{b} \in \mathbb{R}^m$, then we can write expressions like

$$\underbrace{A}_{m\times n}\underbrace{\vec{x}}_{n\times 1} = \underbrace{\vec{b}}_{m\times 1}$$

We will introduce one additional operator on matrices that is useful in this context:

Definition 0.8 (Transpose)**.** The *transpose* of a matrix $A \in \mathbb{R}^{m\times n}$ is a matrix $A^\top \in \mathbb{R}^{n\times m}$ with elements $(A^\top)_{ij} = A_{ji}$.

Example 0.15 (Transposition)**.** The transpose of the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

is given by

$$A^\top = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}.$$

Geometrically, we can think of transposition as flipping a matrix on its diagonal.

This unified treatment of scalars, vectors, and matrices combined with operations like transposition and multiplication can yield slick expressions and derivations of well-known identities. For instance, we can compute the dot products of vectors $\vec{a}, \vec{b} \in \mathbb{R}^n$ via the following sequence of equalities:

$$\vec{a} \cdot \vec{b} = \sum_{k=1}^{n} a_k b_k = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \vec{a}^\top \vec{b}$$

Many identities from linear algebra can be derived by chaining together these operations with a few rules:

$$(A^\top)^\top = A \qquad (A+B)^\top = A^\top + B^\top \qquad (AB)^\top = B^\top A^\top$$

Example 0.16 (Residual norm)**.** Suppose we have a matrix A and two vectors $\vec{x}$ and $\vec{b}$. If we wish to know how well $A\vec{x}$ approximates $\vec{b}$, we might define a *residual* $\vec{r} \equiv \vec{b} - A\vec{x}$; this residual is zero exactly when $A\vec{x} = \vec{b}$. Otherwise, we might use the norm $\|\vec{r}\|_2$ as a proxy for the similarity of $A\vec{x}$ and $\vec{b}$. We can use the identities above to simplify:

$$\begin{aligned}
\|\vec{r}\|_2^2 &= \|\vec{b} - A\vec{x}\|_2^2 \\
&= (\vec{b} - A\vec{x}) \cdot (\vec{b} - A\vec{x}) \text{ as explained in §0.2.3} \\
&= (\vec{b} - A\vec{x})^\top (\vec{b} - A\vec{x}) \text{ by our expression for the dot product above} \\
&= (\vec{b}^\top - \vec{x}^\top A^\top)(\vec{b} - A\vec{x}) \text{ by properties of transposition} \\
&= \vec{b}^\top\vec{b} - \vec{b}^\top A\vec{x} - \vec{x}^\top A^\top \vec{b} + \vec{x}^\top A^\top A\vec{x} \text{ after multiplication}
\end{aligned}$$

```
function MULTIPLY(A, x⃗)
  ▷ Returns b⃗ = Ax⃗, where
  ▷ A ∈ ℝ^{m×n} and x⃗ ∈ ℝ^n
  b⃗ ← 0⃗
  for i ← 1, 2, ..., m
    for j ← 1, 2, ..., n
      b_i ← b_i + a_ij x_j
  return b⃗
```

(a)

```
function MULTIPLY(A, x⃗)
  ▷ Returns b⃗ = Ax⃗, where
  ▷ A ∈ ℝ^{m×n} and x⃗ ∈ ℝ^n
  b⃗ ← 0⃗
  for j ← 1, 2, ..., n
    for i ← 1, 2, ..., m
      b_i ← b_i + a_ij x_j
  return b⃗
```

(b)

FIGURE 0.2 Two implementations of matrix-vector multiplication with different loop ordering.

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

(a)

1	2	3	4	5	6

(b) Row-major

1	3	5	2	4	6

(c) Column-major

FIGURE 0.3 Two possible ways to store a matrix (a) in memory: row-major ordering (b) and column-major ordering (c).

All four terms on the right hand side are scalars, or equivalently 1×1 matrices. Scalars thought of as matrices enjoy one additional nice property $c^\top = c$, since there is nothing to transpose! Thus, we can write

$$\vec{x}^\top A^\top \vec{b} = (\vec{x}^\top A^\top \vec{b})^\top = \vec{b}^\top A\vec{x}$$

This allows us to simplify our expression even more:

$$\begin{aligned} \|\vec{r}\|_2^2 &= \vec{b}^\top \vec{b} - 2\vec{b}^\top A\vec{x} + \vec{x}^\top A^\top A\vec{x} \\ &= \|A\vec{x}\|_2^2 - 2\vec{b}^\top A\vec{x} + \|\vec{b}\|_2^2 \end{aligned}$$

We could have derived this expression using dot product identities, but intermediate steps above will prove useful in our later discussion.

0.3.3 Matrix Storage and Multiplication Methods

Figure 0.2 shows two possible implementations of matrix-vector multiplication. The difference between these two algorithms is subtle and seemingly unimportant: The order of the two loops has been switched. Rounding error aside, these two methods generate the same output and do essentially the same number of arithmetic operations; classical "big-O" analysis from computer science would find these two methods indistinguishable. Surprisingly, however, a number of engineering choices can make one or the other of these choices much faster than the other!

A reasonable model for the memory or RAM in a computer is as a long line of data. For

this reason, we must find ways to "unroll" data from matrix form to something that could be written completely horizontally. Two common patterns are illustrated in Figure 0.3:

- A *row-major* ordering stores the data row-by-row; that is, the first row appears in a contiguous block of memory, then the second, and so on.
- A *column-major* ordering stores the data column-by-column, moving vertically first rather than horizontally.

Now, consider the matrix multiplication method in Figure 0.2(a). This algorithm computes all of b_1 before moving to b_2, b_3, and so on. In doing so, the code moves along the elements of A row-by-row. So, if A is stored in row-major order then the algorithm in Figure 0.2(a) proceeds linearly across its representation in memory (Figure 0.3(b)), whereas if A is stored in column-major order (Figure 0.3(c)) the algorithm effectively jumps around between elements in A. The opposite is true for the algorithm in Figure 0.2(b), which moves linearly through the column-major ordering.

In many hardware implementations, loading data from memory will yield not just the single requested value but also a block of data around the request. The philosophy here is that common algorithms move linearly though data, processing it one element at a time. By pairing e.g. the algorithm in Figure 0.2(a) with the row-major ordering in Figure 0.3(b), we can take advantage of this optimization by moving linearly through the storage of the matrix A; thus, the extra loaded data in the block gets used without having to make an additional request. If we take a nonlinear traversal of the elements of A, this lucky behavior is less likely to happen, which can lead to a significant loss in speed.

For the most part, these engineering considerations are better left for hardware- and implementation-related discussions. Given the huge size of many matrices, vectors, and other data considered in numerical methods, however, they can be key components of high-speed software for these applications.

0.3.4 Model Problem: $A\vec{x} = \vec{b}$

In introductory algebra class, students spend considerable time solving linear systems such as the following for triplets (x, y, z):

$$\begin{aligned} 3x + 2y + 5z &= 0 \\ -4x + 9y - 3z &= -7 \\ 2x - 3y - 3z &= 1 \end{aligned}$$

Our constructions in §0.3.1 allows us to encode such systems in a cleaner fashion:

$$\begin{pmatrix} 3 & 2 & 5 \\ -4 & 9 & -3 \\ 2 & -3 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ -7 \\ 1 \end{pmatrix}$$

More generally, we can write any linear system of equations in the form $A\vec{x} = \vec{b}$ by following the same pattern above; here, the vector $\vec{x}$ is unknown while A and $\vec{b}$ are known. Such a system of equations is *not* always guaranteed to have a solution. For instance, if A contains only zeros, then no $\vec{x}$ will satisfy $A\vec{x} = \vec{b}$ whenever $\vec{b} \neq \vec{0}$. We will defer a general consideration of when a solution exists to our discussion of linear solvers in future chapters.

A key interpretation of the system $A\vec{x} = \vec{b}$ is that it addresses the task:

Write $\vec{b}$ as a linear combination of the columns of A.

Why? Recall from §0.3.1 that the product $A\vec{x}$ encodes a linear combination of the columns of A with weights contained in elements of $\vec{x}$. So, the equation $A\vec{x} = \vec{b}$ sets the linear combination $A\vec{x}$ equal to the given vector $\vec{b}$. Given this interpretation, we define the *column space* of A to be the space of right hand sides $\vec{b}$ for which the system $A\vec{x} = \vec{b}$ has a solution:

Definition 0.9 (Column space and rank)**.** The *column space* of a matrix $A \in \mathbb{R}^{m\times n}$ is the span of the columns of A. It can be written as

$$\text{col } A \equiv \{A\vec{x} : \vec{x} \in \mathbb{R}^n\}.$$

The *rank* of A is the dimension of col A.

One case will dominate our discussion in future chapters. Suppose A is square, so we can write $A \in \mathbb{R}^{n\times n}$. Furthermore, suppose that the system $A\vec{x} = \vec{b}$ has a solution *for all* choices of $\vec{b}$, so by our interpretation above the columns of A must span $\mathbb{R}^n$.

In this case, we can substitute the standard basis $\vec{e}_1, \ldots, \vec{e}_n$ to solve equations of the form $A\vec{x}_i = \vec{e}_i$, yielding vectors $\vec{x}_1, \ldots, \vec{x}_n$. Combining these $\vec{x}_i$'s horizontally into a matrix shows:

$$A\begin{pmatrix} | & | & & | \\ \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ A\vec{x}_1 & A\vec{x}_2 & \cdots & A\vec{x}_n \\ | & | & & | \end{pmatrix}$$
$$= \begin{pmatrix} | & | & & | \\ \vec{e}_1 & \vec{e}_2 & \cdots & \vec{e}_n \\ | & | & & | \end{pmatrix} = I_{n\times n},$$

where $I_{n\times n}$ is the identity matrix from Example 0.11. We will call the matrix with columns $\vec{x}_k$ the *inverse* A^{-1}, which satisfies

$$AA^{-1} = A^{-1}A = I_{n\times n}.$$

By construction, $(A^{-1})^{-1} = A$. If we can find such an inverse, solving any linear system $A\vec{x} = \vec{b}$ reduces to matrix multiplication, since:

$$\vec{x} = I_{n\times n}\vec{x} = (A^{-1}A)\vec{x} = A^{-1}(A\vec{x}) = A^{-1}\vec{b}.$$

0.4 NON-LINEARITY: DIFFERENTIAL CALCULUS

While the beauty and applicability of linear algebra makes it a key target for study, nonlinearities abound in nature, and hence we must design machinery that can deal with this reality.

0.4.1 Differentiation in One Variable

While many functions are *globally* nonlinear, *locally* they exhibit linear behavior. This idea of "local linearity" is one of the main motivators behind differential calculus. For instance, Figure 0.4 shows that if you zoom in close enough to a smooth function eventually it looks like a line. The derivative $f'(x)$ of a function $f(x) : \mathbb{R} \to \mathbb{R}$ is the slope of the approximating line, computed by finding the slope of lines through closer and closer points to x:

$$f'(x) = \lim_{y\to x} \frac{f(y) - f(x)}{y - x}$$

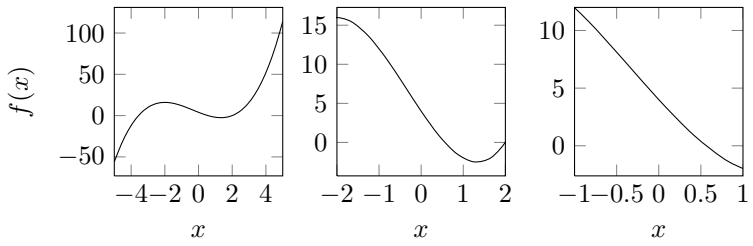


FIGURE 0.4 The closer we zoom into $f(x) = x^3 + x^2 - 8x + 4$, the more it looks like a line.

In reality, taking limits as $y \to x$ may not be possible on a discrete system, so a reasonable question to ask is how well a function $f(x)$ is approximated by a linear approximation through points that are a finite distance apart. We can answer these types of questions using infinitesimal analysis. Take $x, y \in \mathbb{R}$. Then, we can expand:

$$\begin{aligned} f(y) - f(x) &= \int_x^y f'(t)\,dt \text{ by the Fundamental Theorem of Calculus} \\ &= yf'(y) - xf'(x) - \int_x^y tf''(t)\,dt, \text{ after integrating by parts} \\ &= (y-x)f'(x) + y(f'(y) - f'(x)) - \int_x^y tf''(t)\,dt \\ &= (y-x)f'(x) + y\int_x^y f''(t)\,dt - \int_x^y tf''(t)\,dt \\ &\qquad \text{again by the Fundamental Theorem of Calculus} \\ &= (y-x)f'(x) + \int_x^y (y-t)f''(t)\,dt \end{aligned}$$

Rearranging terms and defining $\Delta x \equiv y - x$ shows:

$$\begin{aligned} |f'(x)\Delta x - [f(y) - f(x)]| &= \left| \int_x^y (y-t)f''(t)\,dt \right| \text{ from the relationship above} \\ &\leq |\Delta x| \int_x^y |f''(t)|\,dt, \text{ by the triangle inequality} \\ &\leq D|\Delta x|^2, \text{ assuming } |f''(t)| < D \text{ for some } D > 0. \end{aligned}$$

We can introduce some notation to help express the relationship we have written:

Definition 0.10 (Infinitesimal big-O)**.** We will say $f(x) = O(g(x))$ if there exists a constant $C > 0$ and some $\varepsilon > 0$ such that $|f(x)| \leq C|g(x)|$ for all x with $|x| < \varepsilon$.

This definition is illustrated in Figure 0.5.

Computer scientists may be surprised to see that we are defining "big-O notation" by taking limits as $x \to 0$ rather than $x \to \infty$, but since we are concerned with infinitesimal approximation quality this definition will be more relevant to the discussion at hand.

Our derivation above shows the following key relationship for smooth functions $f : \mathbb{R} \to \mathbb{R}$:

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + O(\Delta x^2)$$

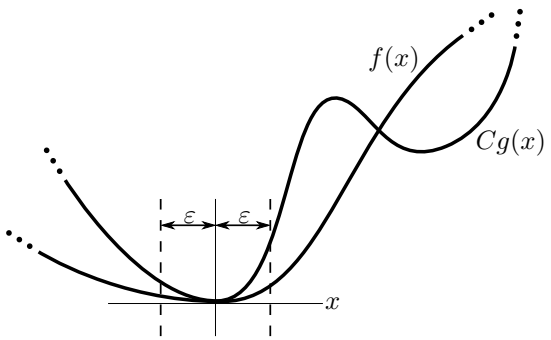


FIGURE 0.5 Big-O notation; in the ε neighborhood of the origin, $f(x)$ is dominated by $Cg(x)$; outside this neighborhood, $Cg(x)$ can dip back down.

This is an instance of Taylor's Theorem, which we will apply copiously when developing strategies for integrating ordinary differential equations. More generally, this theorem shows how to approximate differentiable functions with polynomials:

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + f''(x)\frac{\Delta x^2}{2!} + \cdots + f^{(k)}(x)\frac{\Delta x^k}{k!} + O(\Delta x^{k+1})$$

0.4.2 Differentiation in Multiple Variables

If the function f takes multiple inputs, then it can be written $f(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$ for $\vec{x} \in \mathbb{R}^n$; in other words, to each point $\vec{x} = (x_1, \ldots, x_n)$ in n-dimensional space f assigns a single number $f(x_1, \ldots, x_n)$.

The idea of local linearity must be repaired in this case, because lines are one- rather than n-dimensional objects. Fixing all but one variable, however, brings a return to single-variable calculus. For instance, we could isolate x_1 by writing $g(t) \equiv f(t, x_2, \ldots, x_n)$, where we think of $x_2, \ldots, x_n$ as constants. Then, $g(t)$ is a differentiable function of a single variable that we can differentiate using the machinery in §13.3. We could have substituted t for any of the x_k's, so in general we make the following definition of the *partial derivative* of f:

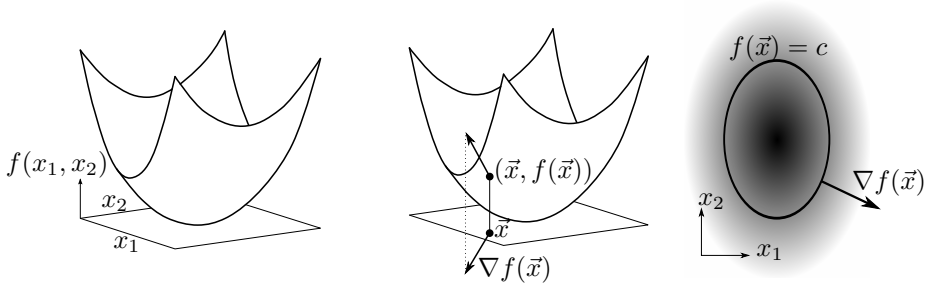
Definition 0.11 (Partial derivative)**.** The k-th *partial derivative* of f, notated $\frac{\partial f}{\partial x_k}$, is given by differentiating f in its k-th input variable:

$$\frac{\partial f}{\partial x_k}(x_1, \ldots, x_n) \equiv \frac{d}{dt} f(x_1, \ldots, x_{k-1}, t, x_{k+1}, \ldots, x_n)|_{t=x_k}$$

The notation "$|_{t=x_k}$" should be read as "evaluated at $t = x_k$."

Example 0.17 (Relativity)**.** The relationship $E = mc^2$ can be thought of as a function mapping pairs (m, c) to a scalar E. Thus, we could write $E(m, c) = mc^2$, yielding the partial derivatives

$$\frac{\partial E}{\partial m} = c^2 \qquad \frac{\partial E}{\partial c} = 2mc$$



(a) Graph of $f(\vec{x})$ (b) Steepest ascent (c) Level sets of $f(\vec{x})$

FIGURE 0.6 (a) We can visualize a function $f(x_1, x_2)$ as a three-dimensional graph; then $\nabla f(\vec{x})$ is the direction on the (x_1, x_2) plane corresponding to the steepest ascent of f on the graph (b). Alternatively, we can think of $f(x_1, x_2)$ as a *color* at (x_1, x_2) (c; dark indicates a low value of f), in which case ∇f points perpendicular to level sets $f(\vec{x}) = c$ in the direction where f is increasing and the image gets lighter.

Using single-variable calculus, for a function $f : \mathbb{R}^n \to \mathbb{R}$ we can write:

$$\begin{aligned} f(\vec{x} + \Delta\vec{x}) &= f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) \\ &= f(x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n) + \frac{\partial f}{\partial x_1}\Delta x_1 + O(\Delta x_1^2) \\ &\qquad \text{by single-variable calculus in the } x_1 \text{ slot} \\ &= f(x_1, \ldots, x_n) + \sum_{k=1}^{n}\left[\frac{\partial f}{\partial x_k}\Delta x_k + O(\Delta x_k^2)\right] \\ &\qquad \text{by repeating this } n-1 \text{ times in the slots } x_2, \ldots, x_n \\ &= f(\vec{x}) + \nabla f(\vec{x}) \cdot \Delta\vec{x} + O(\|\Delta\vec{x}\|^2) \end{aligned}$$

where we define the *gradient* of f as

$$\nabla f(\vec{x}) \equiv \left(\frac{\partial f}{\partial x_1}(\vec{x}), \frac{\partial f}{\partial x_2}(\vec{x}), \cdots, \frac{\partial f}{\partial x_n}(\vec{x})\right) \in \mathbb{R}^n$$

Figure 0.6 illustrates one interpretation of the gradient of a function, which we will reconsider in our discussion of optimization in future chapters. From this relationship, we can differentiate f in any direction $\vec{v}$ by evaluating the corresponding *directional derivative* $D_{\vec{v}}f$:

$$D_{\vec{v}}f(\vec{x}) \equiv \frac{d}{dt}f(\vec{x} + t\vec{v})|_{t=0} = \nabla f(\vec{x}) \cdot \vec{v}.$$

We allow $\vec{v}$ to have any length, with the property $D_{c\vec{v}}f(\vec{x}) = cD_{\vec{v}}f(\vec{x})$.

Example 0.18 ($\mathbb{R}^2$)**.** Take $f(x, y) = x^2y^3$. Then,

$$\frac{\partial f}{\partial x} = 2xy^3 \qquad \frac{\partial f}{\partial y} = 3x^2y^2$$

Equivalently, $\nabla f(x, y) = (2xy^3, 3x^2y^2)$. So, the derivative of f at $(x, y) = (1, 2)$ in the direction $(-1, 4)$ is given by $(-1, 4) \cdot \nabla f(1, 2) = (-1, 4) \cdot (16, 12) = 32$.

In this book, there are a few derivatives that we will use many times. These formulae will appear repeatedly in future chapters and are worth studying independently:

Example 0.19 (Linear functions)**.** It is obvious but worth noting that the gradient of $f(\vec{x}) \equiv \vec{a} \cdot \vec{x} + \vec{c} = (a_1 x_1 + c_1, \ldots, a_n x_n + c_n)$ is $\vec{a}$.

Example 0.20 (Quadratic forms)**.** Take any matrix $A \in \mathbb{R}^{n \times n}$, and define $f(\vec{x}) \equiv \vec{x}^\top A \vec{x}$. Writing this function element-by-element shows

$$f(\vec{x}) = \sum_{ij} A_{ij} x_i x_j.$$

Expanding f and checking this relationship explicitly is worthwhile. Take some $k \in \{1, \ldots, n\}$. Then, we can can separate out all terms containing x_k:

$$f(\vec{x}) = A_{kk} x_k^2 + x_k \left(\sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j \right) + \sum_{i,j \neq k} A_{ij} x_i x_j$$

With this factorization, it is easy to see

$$\frac{\partial f}{\partial x_k} = 2A_{kk} x_k + \left(\sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j \right) = \sum_{i=1}^{n} (A_{ik} + A_{ki}) x_i$$

This sum looks a lot like the definition of matrix-vector multiplication! Combining these partial derivatives into a single vector shows $\nabla f(\vec{x}) = A\vec{x} + A^\top \vec{x}$. In the special case when A is symmetric, that is, when $A^\top = A$, we have the well-known formula $\nabla f(\vec{x}) = 2A\vec{x}$.

We have generalized from $f : \mathbb{R} \to \mathbb{R}$ to $f : \mathbb{R}^n \to \mathbb{R}$. To reach full generality, we should consider $f : \mathbb{R}^n \to \mathbb{R}^m$. In other words, f takes in n numbers and outputs m numbers. Thankfully, this extension is straightforward, because we can think of f as a collection of single-valued functions $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ smashed together into a single vector. Symbolically, we write:

$$f(\vec{x}) = \begin{pmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \\ \vdots \\ f_m(\vec{x}) \end{pmatrix}$$

Each f_k can be differentiated as before, so in the end we get a matrix of partial derivatives called the *Jacobian* of f:

Definition 0.12 (Jacobian)**.** The *Jacobian* of $f : \mathbb{R}^n \to \mathbb{R}^m$ is the matrix $Df(\vec{x}) \in \mathbb{R}^{m \times n}$ with entries

$$(Df)_{ij} \equiv \frac{\partial f_i}{\partial x_j}.$$

Example 0.21 (Jacobian computation)**.** Suppose $f(x, y) = (3x, -xy^2, x + y)$. Then,

$$Df(x, y) = \begin{pmatrix} 3 & 0 \\ -y^2 & -2xy \\ 1 & 1 \end{pmatrix}.$$

Make sure you can derive this computation by hand.

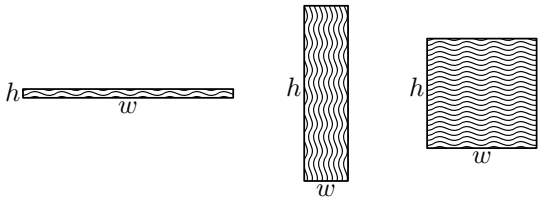


FIGURE 0.7 Three rectangles with the same perimeter $2w + 2h$ but unequal areas wh; the square on the right with $w = h$ maximizes wh over all possible choices with prescribed $2w + 2h = 1$.

Example 0.22 (Matrix multiplication)**.** Unsurprisingly, the Jacobian of $f(\vec{x}) = A\vec{x}$ for matrix A is given by $Df(\vec{x}) = A$.

Here we encounter a common point of confusion. Suppose a function has vector input and scalar output, that is, $f : \mathbb{R}^n \to \mathbb{R}$. We defined the gradient of f as a column vector, so to align this definition with that of the Jacobian we must write $Df = \nabla f^\top$; the transpose allows us to write ∇f as a column vector.

0.4.3 Optimization

From single variable calculus, the minima and maxima of $f : \mathbb{R} \to \mathbb{R}$ must occur at points x satisfying $f'(x) = 0$. This condition is *necessary* rather than *sufficient*: there may exist saddle points x with $f'(x) = 0$ that are not maxima or minima. That said, finding such *critical points* of f can be part of a function minimization algorithm, so long as a subsequent step ensures that the resulting x is actually a minimum/maximum.

If $f : \mathbb{R}^n \to \mathbb{R}$ is minimized or maximized at $\vec{x}$, we have to ensure that there does not exist a single direction Δx from $\vec{x}$ in which f decreases or increases, resp. By the discussion in §13.3, this means we must find points for which $\nabla f = 0$.

Example 0.23 (Critical points)**.** Suppose $f(x, y) = x^2 + 2xy + 4y^2$. Then, $\frac{\partial f}{\partial x} = 2x + 2y$ and $\frac{\partial f}{\partial y} = 2x + 8y$. Thus, critical points of f satisfy:

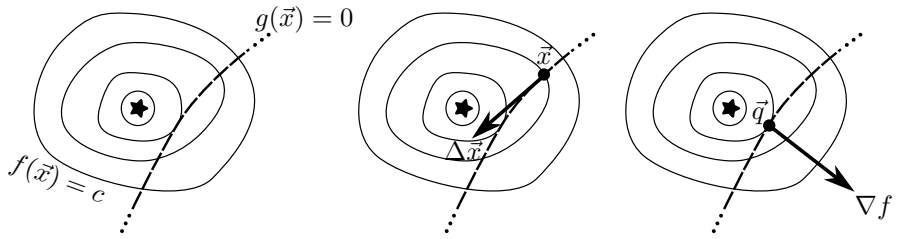
$$2x + 2y = 0 \qquad\qquad 2x + 8y = 0$$

This system is solved by taking $(x, y) = (0, 0)$. Indeed, this is the minimum of f, as can be seen by writing $f(x, y) = (x + y)^2 + 3y^2 \geq 0$.

Example 0.24 (Quadratic functions)**.** Suppose $f(\vec{x}) = \vec{x}^\top A\vec{x} + \vec{b}^\top\vec{x} + c$. Then, from Examples 0.19 and 0.20 we can write $\nabla f(\vec{x}) = (A^\top + A)\vec{x} + \vec{b}$. Thus, critical points $\vec{x}$ of f satisfy $(A^\top + A)\vec{x} + \vec{b} = 0$.

Unlike single-variable calculus, when we do calculus on $\mathbb{R}^n$ we can add *constraints* to our optimization. For now, we will consider the *equality-constrained* case, given by

$$\begin{aligned} &\text{minimize } f(\vec{x}) \\ &\text{such that } g(\vec{x}) = \vec{0}. \end{aligned}$$



(a) Constrained optimization (b) Suboptimal $\vec{x}$ (c) Optimal $\vec{q}$

FIGURE 0.8 (a) An equality-constrained optimization. Without constraints, $f(\vec{x})$ is minimized at the star; solid lines show isocontours $f(\vec{x}) = c$ for increasing c. Minimizing $f(\vec{x})$ subject to $g(\vec{x}) = 0$ forces $\vec{x}$ to be on the dashed curve. (b) The point $\vec{x}$ is suboptimal since moving in the $\Delta\vec{x}$ direction decreases $f(\vec{x})$ while maintaining $g(\vec{x}) = 0$. (c) The point $\vec{q}$ is optimal since decreasing f from $f(\vec{q})$ would require moving in the $-\nabla f$ direction, which is perpendicular to the curve $g(\vec{x}) = 0$.

Example 0.25 (Rectangle areas)**.** Suppose a rectangle has width w and height h. A classic geometry problem is to maximize area with a fixed perimeter 1:

$$\begin{aligned} &\text{maximize } wh \\ &\text{such that } 2w + 2h - 1 = 0 \end{aligned}$$

This problem is illustrated in Figure 0.7.

When we add this constraint, we can no longer expect that minima or maxima satisfy $\nabla f(\vec{x}) = 0$, since these points might not satisfy $g(\vec{x}) = 0$.

For now, suppose $g : \mathbb{R}^n \to \mathbb{R}$, so we only have one equality constraint; an example for $n = 2$ is shown in Figure 0.8. We define the set of points satisfying the equality constraint as $S_0 \equiv \{\vec{x} : g(\vec{x}) = 0\}$. Any two $\vec{x}, \vec{y} \in S_0$ satisfy the relationship $g(\vec{y}) - g(\vec{x}) = 0 - 0 = 0$. Applying Taylor's Theorem, if $\vec{y} = \vec{x} + \Delta\vec{x}$ for small $\Delta\vec{x}$, then

$$g(\vec{y}) - g(\vec{x}) = \nabla g(\vec{x}) \cdot \Delta\vec{x} + O(\|\Delta\vec{x}\|^2).$$

In other words, if $g(\vec{x}) = 0$ and $\nabla g(\vec{x}) \cdot \Delta\vec{x} = 0$ then $g(\vec{x} + \Delta\vec{x}) \approx 0$.

If $\vec{x}$ is a minimum of the constrained optimization problem above, then any small displacement $\vec{x}$ to $\vec{x} + \vec{v}$ should cause an increase from $f(\vec{x})$ to $f(\vec{x} + \vec{v})$. Since we only care about displacements $\vec{v}$ preserving the $g(\vec{x} + \vec{v}) = c$ constraint, from our argument above we want $\nabla f \cdot \vec{v} = 0$ for all $\vec{v}$ satisfying $\nabla g(\vec{x}) \cdot \vec{v} = 0$. In other words, ∇f and ∇g must be parallel, a condition we can write as $\nabla f = \lambda \nabla g$ for some $\lambda \in \mathbb{R}$, illustrated in Figure 0.8(c).

Define

$$\Lambda(\vec{x}, \lambda) \equiv f(\vec{x}) - \lambda g(\vec{x}).$$

Then, critical points of Λ without constraints satisfy:

$$\begin{aligned} \frac{\partial \Lambda}{\partial \lambda} &= -g(\vec{x}) = 0 \text{ by the constraint } g(\vec{x}) = 0 \\ \nabla_{\vec{x}} \Lambda &= \nabla f(\vec{x}) - \lambda \nabla g(\vec{x}) = 0, \text{ as argued above} \end{aligned}$$

In other words, critical points of Λ with respect to both λ and $\vec{x}$ satisfy $g(\vec{x}) = 0$ and $\nabla f(\vec{x}) = \lambda \nabla g(\vec{x})$, exactly the optimality conditions we derived!

More generally, extending our argument to $g : \mathbb{R}^n \to \mathbb{R}^k$ yields the following theorem:

Theorem 0.1 (Method of Lagrange multipliers)**.** Critical points of the equality-constrained optimization problem above are (unconstrained) critical points of the Lagrange multiplier function

$$\Lambda(\vec{x}, \vec{\lambda}) \equiv f(\vec{x}) - \vec{\lambda} \cdot g(\vec{x}),$$

with respect to both $\vec{x}$ and $\vec{\lambda}$.

This theorem provides an analog of the condition $\nabla f(\vec{x}) = \vec{0}$ when equality constraints $g(\vec{x}) = \vec{0}$ are added to an optimization. It is a cornerstone of *variational* methods we will consider, where we seek minima or maxima of an energy function to solve a given numerical problem. We conclude with a number of examples applying this theorem; understanding these examples is crucial to our development of numerical methods in future chapters.

Example 0.26 (Maximizing area)**.** Continuing Example 0.25, we define the Lagrange multiplier function $\Lambda(w, h, \lambda) = wh - \lambda(2w + 2h - 1)$. Differentiating Λ with respect to w, h, and λ provides the following optimality conditions:

$$0 = \frac{\partial \Lambda}{\partial w} = h - 2\lambda \qquad 0 = \frac{\partial \Lambda}{\partial h} = w - 2\lambda \qquad 0 = \frac{\partial \Lambda}{\partial \lambda} = 1 - 2w - 2h$$

So, critical points of the area wh under the constraint $2w + 2h = 1$ satisfy

$$\begin{pmatrix} 0 & 1 & -2 \\ 1 & 0 & -2 \\ 2 & 2 & 0 \end{pmatrix} \begin{pmatrix} w \\ h \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Solving the system shows $w = h = 1/4$ (and $\lambda = 1/8$). In other words, for a fixed amount of perimeter, the rectangle with maximal area is a square.

Example 0.27 (Eigenproblems)**.** Suppose that A is a symmetric positive definite matrix, meaning $A^\top = A$ (symmetric) and $\vec{x}^\top A\vec{x} > 0$ for all $\vec{x} \in \mathbb{R}^n \backslash \{\vec{0}\}$ (positive definite). We may wish to minimize $\vec{x}^\top A\vec{x}$ subject to $\|x\|_2^2 = 1$ for a given matrix $A \in \mathbb{R}^{n \times n}$; that without the constraint the function is minimized at $\vec{x} = \vec{0}$. We define the Lagrange multiplier function

$$\Lambda(\vec{x}, \lambda) = \vec{x}^\top A\vec{x} - \lambda(\|\vec{x}\|_2^2 - 1) = \vec{x}^\top A\vec{x} - \lambda(\vec{x}^\top \vec{x} - 1).$$

Differentiating with respect to $\vec{x}$, we find $0 = \nabla_{\vec{x}}\Lambda = 2A\vec{x} - 2\lambda\vec{x}$. In other words, critical points of $\vec{x}$ are exactly the *eigenvectors* of the matrix A:

$$A\vec{x} = \lambda\vec{x}, \text{ with } \|x\|_2^2 = 1.$$

At these critical points, we can evaluate the objective function as $\vec{x}^\top A\vec{x} = \vec{x}^\top \lambda\vec{x} = \lambda\|\vec{x}\|_2^2 = \lambda$. Hence, the minimizer of $\vec{x}^\top A\vec{x}$ subject $\|\vec{x}\|_2^2 = 1$ is the eigenvector $\vec{x}$ with minimum eigenvalue λ; we will provide practical applications and solution techniques for this optimization problem in detail in Chapter 5.

0.5 EXERCISES

0.1[SC] Illustrate the gradient of $f(x, y) = x^2 + y^2$ and $g(x, y) = \sqrt{x^2 + y^2}$ on the plane, and show that $\|\nabla g(x, y)\|_2$ is constant away from the origin.

0.2[DH] Compute the dimensions of each of the following sets:

(a) $\text{col}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$.

(b) $\text{span}\{(1,1,1),(1,-1,1),(-1,1,1),(1,1,-1)\}$

(c) $\text{span}\{(2,7,9),(3,5,1),(0,1,0)\}$

(d) $\text{col}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

0.3 Suppose $A, B \in \mathbb{R}^{n\times n}$ and $\vec{a}, \vec{b} \in \mathbb{R}^n$. Find a linear system of equations satisfied by any $\vec{x}$ minimizing the energy $\|A\vec{x} - \vec{a}\|_2^2 + \|B\vec{x} - \vec{b}\|_2^2$.

0.4 Take $C^1(\mathbb{R})$ to be the set of continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$. Why is $C^1(\mathbb{R})$ a vector space? Show that $C^1(\mathbb{R})$ has dimension ∞.

0.5 Suppose the rows of $A \in \mathbb{R}^{m\times n}$ are given by the transposes of $\vec{r}_1, \ldots, \vec{r}_m \in \mathbb{R}^n$ and the columns of $A \in \mathbb{R}^{m\times n}$ are given by $\vec{c}_1, \ldots, \vec{c}_n \in \mathbb{R}^m$. That is,

$$A = \begin{pmatrix} - & \vec{r}_1^\top & - \\ - & \vec{r}_2^\top & - \\ & \vdots & \\ - & \vec{r}_m^\top & - \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \vec{c}_1 & \vec{c}_2 & \cdots & \vec{c}_n \\ | & | & & | \end{pmatrix}.$$

Give expressions for the elements of $A^\top A$ and $AA^\top$ in terms of these vectors.

0.6 Give a linear system of equations satisfied by minima of the energy $f(\vec{x}) = \|A\vec{x} - \vec{b}\|_2$ with respect to $\vec{x}$, for $\vec{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $\vec{b} \in \mathbb{R}^m$.

0.7 Suppose $A, B \in \mathbb{R}^{n\times n}$. Formulate a condition for vectors $\vec{x} \in \mathbb{R}^n$ to be critical points of $\|A\vec{x}\|_2$ subject to $\|B\vec{x}\|_2 = 1$. Also, give an alternative expression for the optimal values of $\|A\vec{x}\|_2$.

0.8 Fix some vector $\vec{a} \in \mathbb{R}^n \backslash \{\vec{0}\}$ and define $f(\vec{x}) = \vec{a} \cdot \vec{x}$. Give an expression for the maximum of $f(\vec{x})$ subject to $\|\vec{x}\|_2 = 1$.

0.9 Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric, and define the *Rayleigh quotient* function $R(\vec{x})$ as

$$R(\vec{x}) \equiv \frac{\vec{x}^\top A\vec{x}}{\|\vec{x}\|_2}.$$

Show that minimizers of $R(\vec{x})$ subject to $\vec{x} \neq \vec{0}$ are eigenvectors of A.

0.10 Show that $(A^\top)^{-1} = (A^{-1})^\top$ when $A \in \mathbb{R}^{n\times n}$ is invertible. If $B \in \mathbb{R}^{n\times n}$ is also invertible, show $(AB)^{-1} = B^{-1}A^{-1}$.

0.11 Suppose $A(t)$ is a function taking a parameter t and returning an invertible square matrix $A(t) \in \mathbb{R}^{n\times n}$; we can write $A : \mathbb{R} \to \mathbb{R}^{n\times n}$. Assuming each element $a_{ij}(t)$ of

$A(t)$ is a differentiable function of t, define the derivative matrix $\frac{dA}{dt}(t)$ as the matrix whose elements are $\frac{da_{ij}}{dt}(t)$. Then, verify the following identity:

$$\frac{d(A^{-1})}{dt} = -A^{-1}\frac{dA}{dt}A^{-1}.$$

Hint: Start from the identity $A^{-1}(t) \cdot A(t) = I_{n\times n}$.

0.12 A matrix $A \in \mathbb{R}^{n\times n}$ is *idempotent* if it satisfies $A^2 = A$.

(a) Suppose $B \in \mathbb{R}^{m\times k}$ is constructed so that $B^\top B$ is invertible. Show that the matrix $B(B^\top B)^{-1}B^\top$ is idempotent.

(b) If A is idempotent, show that $I_{n\times n} - A$ is also idempotent.

(c) If A is idempotent, show that $\frac{1}{2}I_{n\times n} - A$ is invertible and give an expression for its inverse.

(d) Suppose A is idempotent and that we are given $\vec{x} \neq \vec{0}$ and $\lambda \in \mathbb{R}$ satisfying $A\vec{x} = \lambda\vec{x}$. Show that $\lambda \in \{0, 1\}$.

CHAPTER 1

Numerics and Error Analysis

CONTENTS

1.1 Storing Numbers with Fractional Parts 25
1.1.1 Fixed Point Representations 26
1.1.2 Floating Point Representations 27
1.1.3 More Exotic Options 29
1.2 Understanding Error 30
1.2.1 Classifying Error 31
1.2.2 Conditioning, Stability, and Accuracy 32
1.3 Practical Aspects 34
1.3.1 Computing Vector Norms 34
1.3.2 Larger-Scale Example: Summation 35

NUMERICAL analysis introduces a shift from working with `int`s and `long`s to `float`s and `double`s. This seemingly innocent transition shatters intuition from integer arithmetic, requiring adjustment of how we must think about basic algorithmic design and implementation. Unlike discrete algorithms, numerical algorithms cannot always yield exact solutions even to well-studied and well-posed problems. Operation counting no longer reigns supreme; instead, even basic techniques require careful analysis of the trade-offs between timing, approximation error, and other considerations. In this chapter, we will explore the typical factors affecting the quality of a numerical algorithm; these factors set numerical algorithms apart from their discrete counterparts.

1.1 STORING NUMBERS WITH FRACTIONAL PARTS

Most computers store data in *binary* format. In binary, integers are decomposed into powers of two. For instance, we can convert 463 to binary using the following table:

1	1	1	0	0	1	1	1	1
2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0

This table illustrates the fact that 463 has a unique decomposition into powers of two as:

$$\begin{aligned} 463 &= 256 + 128 + 64 + 8 + 4 + 2 + 1 \\ &= 2^8 + 2^7 + 2^6 + 2^3 + 2^2 + 2^1 + 2^0 \end{aligned}$$

Issues of overflow aside, all positive integers can be written in this form. Negative numbers also can be represented either by introducing a leading sign bit (e.g. 1 for "positive" and 0 for "negative") or by using a "two's complement" trick.

The binary system admits an extension to numbers with fractional parts by including *negative* powers of two. For instance, 463.25 can be decomposed by adding two slots:

1	1	1	0	0	1	1	1	1.	0	1
2^8	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0	2^{-1}	2^{-2}

Representing fractional parts of numbers this way, however, is not nearly as well-behaved as representing integers. For instance, writing the fraction $1/3$ in binary requires infinitely many digits:

$$\frac{1}{3} = 0.0101010101\cdots_2$$

More generally, there exist numbers at all scales that cannot be represented using a finite binary string. In fact, all irrational numbers, like $\pi = 11.00100100001\ldots_2$, have infinitely-long expansions regardless of which (integer) base you use!

Since computers have finite amount storage capacity, systems processing values in $\mathbb{R}$ instead of $\mathbb{Z}$ are forced to approximate or restrict values that can be processed. This leads to many points of confusion while coding, as in the following snippet of C++ code:

```
double x = 1.0;
double y = x / 3.0;
if (x == y*3.0) cout << "They are equal!";
else cout << "They are NOT equal.";
```

Contrary to intuition, this program prints `"They are NOT equal."` Why? Since $1/3$ cannot be written as a finite-length binary string, the definition of `y` makes an approximation, rounding to the nearest number representable in the `double` data type. Thus, `y*3.0` is *close to but not exactly* 3. One way to fix this issue is to allow for some tolerance:

```
double x = 1.0;
double y = x / 3.0;
if (fabs(x-y*3.0) < numeric_limits<double>::epsilon)
     cout << "They are equal!";
else cout << "They are NOT equal.";
```

Here, we check that `x` and `y*3.0` are near enough to each other to be reasonably considered identical rather than whether they are exactly equal. The tolerance `epsilon` expresses how far apart values should be before we are are confident they are different. It may need to be adjusted depending on context. This example raises a crucial point:

Rarely if ever should the operator == and its equivalents be used on fractional values. Instead, some *tolerance* should be used to check if numbers are equal.

There is a trade-off here: the size of the tolerance defines a line between equality and "close-but-not-the-same," which must be chosen carefully for a given application.

The error generated by a numerical algorithm depends on the choice of *representations* for real numbers. Each representation has its own compromise between speed, accuracy, range of representable values, and so on. Keeping the example above and its resolution in mind, we now consider a few options for representing numbers discretely.

1.1.1 Fixed Point Representations

The most straightforward way to store fractions is to use a *fixed* decimal point. That is, as in the example above, we represent values by storing 0-or-1 coefficients in front of powers of two that range from 2^{-k} to 2^{ℓ} for some $k, \ell \in \mathbb{Z}$. For instance, representing all nonnegative values between 0 and 127.75 in increments of $1/4$ is as easy as taking $k = 2$ and $\ell = 7$; in this situation, we represent these values using 9 binary digits, of which two occur after the decimal point.

The primary advantage of this representation is that many arithmetic operations can be carried out using the same machinery already in place for integers. For example, if a and b are written in fixed-point format, we can write:

$$a + b = (a \cdot 2^k + b \cdot 2^k) \cdot 2^{-k}.$$

The values $a \cdot 2^k$ and $b \cdot 2^k$ are integers, the summation on the right hand side is an integer operation. This observation essentially shows that fixed-point addition can be carried out using integer addition essentially by "ignoring" the decimal point. In this way, rather than using specialized hardware, the preexisting integer arithmetic logic unit (ALU) can carry out fixed-point mathematics quickly.

Fixed-point arithmetic may be fast, but it suffers from serious precision issues. In particular, it is often the case that the output of a binary operation like multiplication or division can require more bits than the operands. For instance, suppose we include one decimal point of precision and wish to carry out the product $1/2 \cdot 1/2 = 1/4$. We write $0.1_2 \times 0.1_2 = 0.01_2$, which gets truncated to 0. More broadly, it is straightforward to combine fixed point numbers in a reasonable way and get an unreasonable result.

Due to these drawbacks, most major programming languages do not by default include a fixed-point data type. The speed and regularity of fixed-pont arithmetic, however, can be a considerable advantage for systems that favor timing over accuracy. Some lower-end graphics processing units (GPU) implement only fixed-point operations since a few decimal points of precision is sufficient for many graphical applications.

1.1.2 Floating Point Representations

One of many numerical challenges in scientific computing is the extreme range of scales that can appear. For example, chemists deal with values anywhere between 9.11×10^{-31} (the mass of an electron in kilograms) and 6.022×10^{23} (the Avogadro constant). An operation as innocent as a change of units can cause a sudden transition between scales: The same observation written in kilograms per lightyear will look considerably different in megatons per mile. As numerical analysts, we are charged with writing software that can transition gracefully between these scales without imposing unnatural restrictions on the client.

Scientists deal with similar issues when recording experimental measurements, and their methods can motivate our formats for storing real numbers on a computer. Most prominently, one of the following representations is more compact than the other:

$$6.022 \times 10^{23} = 602,200,000,000,000,000,000,000$$

Not only does the representation on the left avoid writing an unreasonable number of zeros, but it also reflects the fact that we may not know Avogadro's constant beyond the second 2.

Similarly, in the absence of exceptional scientific equipment, the difference between 6.022×10^{23} and $6.022 \times 10^{23} + 9.11 \times 10^{-31}$ likely is negligible, in the sense that this tiny perturbation is dwarfed by the error of truncating 6.022 to three decimal points. More formally, we say that 6.022×10^{23} has only four *digits of precision* and probably represents some range of possible measurements $[6.022 \times 10^{23} - \varepsilon, 6.022 \times 10^{23} + \varepsilon]$ for some $\varepsilon \approx 0.001 \times 10^{23}$.

Our first observation allwed us to shorten the representation of 6.022×10^{23} by writing it in *scientific notation*. This number system separates the "interesting" digits of a number from its order of magnitude by writing it in the form $a \times 10^e$ for some $a \sim 1$ and $e \in \mathbb{Z}$. We call this format the *floating-point* form of a number, because unlike the fixed-point setup in

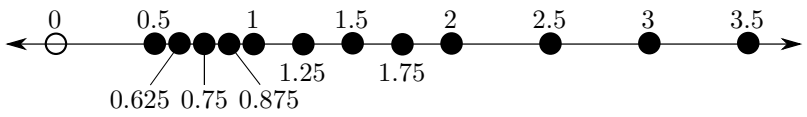


FIGURE 1.1 The values from Example 1.1 plotted on a number line; typical for floating-point number systems, they are unevenly spaced between the minimum (0.5) and the maximum (3.5).

§1.1.1, the decimal point "floats" so that a is on a reasonable scale. Usually a is called the *significand* and e is called the *exponent*.

Floating-point systems are defined using three parameters:

- The *base* or *radix* $b \in \mathbb{N}$. For scientific notation explained above, the base is $b = 10$; for binary systems the basis is $b = 2$.
- The *precision* $p \in \mathbb{N}$ representing the number of digits used to store the significand.
- The range of exponents $[L, U]$ representing the allowable values for b.

The expansion looks like:

$$\underbrace{\pm}_{\text{sign}} \underbrace{(d_0 + d_1 \cdot b^{-1} + d_2 \cdot b^{-2} + \cdots + d_{p-1} \cdot b^{1-p})}_{\text{significand}} \times \underbrace{b^e}_{\text{exponent}}$$

where each digit d_k is in the range $[0, b-1]$ and $e \in [L, U]$. When $b = 2$, an extra bit of precision can be gained by *normalizing* floating point values and assuming the most significant digit d_0 is one; this change, however, requires special treatment of the value 0.

Floating point representations have a curious property that can affect software in unexpected ways: Their spacing is uneven. For example, the number of values representable between b and b^2 is the same as that between b^2 and b^3 even though usually $b^3 - b^2 > b^2 - b$. To understand the precision possible with a given number system, we will define the *machine precision* ε_m as the smallest $\varepsilon_m > 0$ such that $1 + \varepsilon_m$ is representable. Then, numbers like $b + \varepsilon_m$ are not expressible in the number system because ε_m is too small!

Example 1.1 (Floating point)**.** Suppose we choose $b = 2$, $L = -1$, and $U = 1$. If we choose to use three digits of precision, we might choose to write numbers in the form

$$1.\square\square \times 2^{\square}.$$

Notice this number system does not include 0. The possible significands are $1.00_2 = 1_{10}$, $1.01_2 = 1.25_{10}$, $1.10_2 = 1.5_{10}$, and $1.11_2 = 1.75_{10}$. Since $L = -1$ and $U = 1$, these significands can be scaled by $2^{-1} = 0.5_{10}$, $2^0 = 1_{10}$, and $2^1 = 2_{10}$. With this information in hand, we can list all the possible values in our number system:

Significand	$\times 2^{-1}$	$\times 2^0$	$\times 2^1$
1.00_{10}	0.500_{10}	1.000_{10}	2.000_{10}
1.25_{10}	0.625_{10}	1.250_{10}	2.500_{10}
1.50_{10}	0.750_{10}	1.500_{10}	3.000_{10}
1.75_{10}	0.875_{10}	1.750_{10}	3.500_{10}

These values are plotted in Figure 1.1; as expected, they are unevenly spaced and bunch

toward zero. Also, notice the gap between 0 and 0.5 in this sampling of values; some number systems introduce evenly-spaced *subnormal* values to fill in this gap, albeit with less precision. Machine precision for this number system is $\varepsilon_m = 0.25$, the smallest displacement possible above 1.

By far the most common format for storing floating point numbers is provided by the IEEE 754 standard. This standard specifies several classes of floating point numbers. For instance, a double-precision floating point number is written in base $b = 2$ (as are all numbers in this format), with a single $\pm$ sign bit, 52 digits for d, and a range of exponents between -1022 and 1023. The standard also specifies how to store $\pm\infty$ and values like NaN, or "not-a-number," reserved for the results of computations like $10/0$.

IEEE 754 also includes agreed-upon conventions for rounding when an operation results in a number not represented in the standard. For instance, a common unbiased strategy for rounding computations is *round to nearest, ties to even,* which breaks equidistant ties by rounding to the nearest floating point value with an even least-significant (rightmost) bit. There are many equally legitimate strategies for rounding; choosing a single one guarantees that scientific software will work identically on all client machines regardless of their particular processor or compiler.

1.1.3 More Exotic Options

For most of this book, we will assume that fractional values are stored in floating-point format unless otherwise noted. This, however, is not to say that other numerical systems do not exist, and for specific applications an alternative choice might be necessary. We acknowledge some of those situations here.

The headache of adding tolerances to account for rounding errors might be unacceptable for some applications. This situation appears in computational geometry, e.g. when the difference between *nearly-* and *completely*-parallel lines may be a difficult distinction to make. One solution might be to use *arbitrary-precision arithmetic,* that is, to implement fractional arithmetic without rounding or error of any sort.

Arbitrary-precision arithmetic requires a specialized implementation and careful consideration for what types of values you need to represent. For instance, it might be the case that rational numbers $\mathbb{Q}$ are sufficient for a given application, which can be written as ratios a/b for $a, b \in \mathbb{Z}$. Basic arithmetic operations can be carried out in $\mathbb{Q}$ without any loss in precision. For instance, it is easy to see

$$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd} \qquad\qquad \frac{a}{b} \div \frac{c}{d} = \frac{ad}{bc}.$$

Arithmetic in the rationals precludes the existence of a square root operator, since values like $\sqrt{2}$ are irrational. Also, this representation is nonunique, since e.g. $a/b = 5a/5b$, and thus certain operations may require additional routines for simplifying fractions.

In other situations it may be useful to bracket error by representing values alongside error estimates as a pair $a, \varepsilon \in \mathbb{R}$; we think of the pair (a, ε) as the range $a \pm \varepsilon$. Then, arithmetic operations also update not only the value but also the error estimate, as in

$$(x \pm \varepsilon_1) + (y \pm \varepsilon_2) = (x + y) \pm (\varepsilon_1 + \varepsilon_2 + \text{error}(x + y)),$$

where the final term represents an estimate of the error induced by adding x and y. Maintaining error bars in this fashion keeps track of confidence in a given value, which can be informative for scientific calculations.

1.2 UNDERSTANDING ERROR

With the exception of the arbitrary-precision systems described in §1.1.3, nearly every computerized representation of real numbers with fractional parts is forced to employ rounding and other approximations. Rounding, however, represents one of many sources of error typically encountered in numerical systems:

- *Truncation* error comes from rounding and other approximations used to deal with the fact that we can only represent a finite set of values using most computational number systems. For example, it is impossible to write π exactly as an IEEE 754 floating point value, so in practice its value is truncated after a finite number of digits.
- *Discretization* error comes from our computerized adaptations of calculus, physics, and other aspects of continuous mathematics. For instance, a numerical system might attempt to approximate the derivative of a function $f(t)$ using *divided differences*:

  $$f'(t) \approx \frac{f(t+\varepsilon) - f(t)}{\varepsilon}$$

  for some fixed choice of $\varepsilon > 0$. This approximation is a legitimate and useful one, but since we must use a finite $\varepsilon > 0$ rather than taking a limit as $\varepsilon \to 0$, the resulting value for $f'(t)$ is only accurate to some number of digits.
- *Modeling* error comes from having incomplete or inaccurate descriptions of the problems we wish to solve. For instance, a simulation predicting weather in Germany may choose to neglect the collective flapping of butterfly wings in Malaysia, although the displacement of air by these butterflies might perturb the weather patterns elsewhere. Furthermore, constants such as the speed of light or acceleration due to gravity might be provided to the system with a limited degree of accuracy.
- *Input* error can come from user-generated approximations of parameters of a given system (and from typos!). Simulation and numerical techniques can help answer "what if" questions, in which exploratory choices of input setups are chosen just to get some idea of how a system behaves. In this case, a highly-accurate simulation might be a waste of computational time, since the inputs to the simulation were so rough.

Example 1.2 (Computational physics)**.** Suppose we are designing a system for simulating planets as they revolve around the sun. The system essentially solves Newton's equation $F = ma$ by integrating forces forward in time. Examples of error sources in this system might include:

- *Truncation error:* Truncating the product ma to IEEE floating point precision
- *Discretization error:* Using divided differences as above to approximate the velocity and acceleration of each planet
- *Modeling error:* Neglecting to simulate the moon's effects on the earth's motion within the planetary system
- *Input error:* Evaluating the cost of sending garbage into space rather than risking a Wall-E style accumulation on Earth, but only guessing the total amount of garbage to jettison monthly

1.2.1 Classifying Error

Given our previous discussion, the following two numbers might be regarded as having the same amount of error:

$$1 \pm 0.01$$
$$10^5 \pm 0.01$$

Both $[1 - 0.01, 1 + 0.01]$ and $[10^5 - 0.01, 10^5 + 0.01]$ have the same width, but the latter appears to encode a more confident measurement because the error 0.01 is much smaller *relative* to 10^5 than to 1.

The distinction between these two classes of error is described by distinguishing between *absolute* error and *relative* error:

Definition 1.1 (Absolute error)**.** The *absolute error* of a measurement is the difference between the approximate value and its underlying true value.

Definition 1.2 (Relative error)**.** The *relative error* of a measurement is the absolute error divided by the true value.

Absolute error is measured in input units, while relative error is measured as a percentage.

Example 1.3 (Absolute and relative error)**.** Absolute and relative error can be used to express uncertainty in a measurement as follows:

$$\text{Absolute: } 2 \text{ in} \pm 0.02 \text{ in}$$
$$\text{Relative: } 2 \text{ in} \pm 1\%$$

Example 1.4 (Catastrophic cancellation)**.** Suppose we wish to compute the difference $d \equiv 1 - 0.99 = 0.01$. Thanks to an inaccurate representation, we may only know these two values up to ± 0.004. Assuming that we can carry out the subtraction step without error, we are left with the following expression for absolute error:

$$d = 0.01 \pm 0.008$$

In other words, we know d is somewhere in the range $[0.002, 0.018]$. From an absolute perspective, this error may be fairly small. Suppose we attempt to calculate relative error:

$$\frac{|0.002 - 0.01|}{0.01} = \frac{|0.018 - 0.01|}{0.01} = 80\%$$

Thus, although 1 and 0.99 are known with relatively small error, the difference has enormous relative error of 80%. This phenomenon, known as *catastrophic cancellation*, is a danger associated with subtracting two nearby values, yielding a result close to zero.

In most applications the *true* value is unknown; after all, if it were known, there would be no need for an approximation in the first place. Hence it is difficult to compute relative error in closed form. One possible resolution is to be conservative when carrying out computations: At each step take the largest possible error estimate and propagate these estimates forward as necessary. Such conservative estimates are powerful in that when they are small we can be very confident in our output.

An alternative resolution is to acknowledge *what* you can measure; this resolution requires somewhat more intricate arguments but will appear as a theme in future chapters. For

instance, suppose we wish to solve the equation $f(x) = 0$ for x given a function $f : \mathbb{R} \to \mathbb{R}$. Our computational system may yield some x_{est} satisfying $f(x_{\text{est}}) = \varepsilon$ for some ε with $|\varepsilon| \ll 1$. If x_0 is the true root satisfying $f(x_0) = 0$, we may not be able to evaluate the difference $|x_0 - x_{\text{est}}|$ since x_0 is unknown. On the other hand, by evaluating f we can compute $|f(x_{\text{est}}) - f(x_0)| \equiv |f(x_{\text{est}})|$ since $f(x_0) = 0$ by definition. This difference of f values gives a proxy for error that still is zero exactly when $x_{\text{est}} = x_0$.

This example illustrates the distinction between *forward* and *backward* error. Forward error is the most direct definition of error as the difference between the approximated and actual solution, but as we have discussed it is not always computable. Contrastingly, backward error is a calculable proxy for error *correlated* with forward error. We can adjust the definition and interpretation of backward error as we consider different problems, but one suitable—if vague—definition is as follows:

Definition 1.3 (Backward error)**.** The *backward error* of an approximate solution to a numerical problem is the amount the problem statement would have to change to make the approximate solution exact.

This definition is somewhat obtuse, so we illustrate its application to a few scenarios.

Example 1.5 (Linear systems)**.** Suppose we wish to solve the $n \times n$ linear system $A\vec{x} = \vec{b}$. Label the true solution as $\vec{x}_0 \equiv A^{-1}\vec{b}$. In reality, due to truncation error and other issues, our system yields a near-solution $\vec{x}_{\text{est}}$. The forward error of this approximation is the difference $\vec{x}_{\text{est}} - \vec{x}_0$; in practice this difference is impossible to compute since we do not know $\vec{x}_0$. In reality, $\vec{x}_{\text{est}}$ is the *exact* solution to a modified system $A\vec{x} = \vec{b}_{\text{est}}$ for $\vec{b}_{\text{est}} \equiv A\vec{x}_{\text{est}}$; thus, we might measure backward error in terms of the difference $\vec{b} - \vec{b}_{\text{est}}$. Unlike the forward error, this error is easily computable without inverting A, and it is easy to see that $\vec{x}_{\text{est}}$ is a solution to the problem exactly when backward (or forward) error is zero.

Example 1.6 (Solving equations, from [54] example 1.5)**.** Suppose we write a function for finding square roots of positive numbers that outputs $\sqrt{2} \approx 1.4$. The forward error is $|1.4 - \sqrt{2}| \approx 0.0142$. The backward error is $|1.4^2 - 2| = 0.04$.

These examples demonstrate that backward error can be much easier to compute than forward error. For example, evaluating forward error in Example 1.5 required inverting a matrix A while evaluating backward error required only multiplication by A. Similarly, in Example 1.6 transitioning from forward error to backward error replaced square root computation with multiplication.

1.2.2 Conditioning, Stability, and Accuracy

In nearly any numerical problem, zero backward error implies zero forward error and vice versa. Thus, a piece of software designed to solve such a problem surely can terminate if it finds that a candidate solution has zero backward error. But what if backward error is small but nonzero? Does this condition necessarily imply small forward error? We must address such questions to justify replacing forward error with backward error for evaluating the success of a numerical algorithm.

The relationship between forward and backward error can be different for each problem we wish to solve, so in the end we make the following rough classification:

- A problem is *insensitive* or *well-conditioned* when small amounts of backward error imply small amounts of forward error. In other words, a small perturbation to the

statement of a well-conditioned problem yields only a small perturbation of the true solution.

- A problem is *sensitive* or *poorly-conditioned* when this is not the case.

Example 1.7 ($ax = b$)**.** Suppose as a toy example that we want to find the solution $x_0 \equiv {}^{b}/_{a}$ to the linear equation $ax = b$ for $a, x, b \in \mathbb{R}$. Forward error of a potential solution x is given by $|x - x_0|$ while backward error is given by $|b - ax| = |a(x - x_0)|$. So, when $|a| \gg 1$, the problem is well-conditioned since small values of backward error $a(x - x_0)$ imply even smaller values of $x - x_0$; contrastingly, when $|a| \ll 1$ the problem is ill-conditioned, since even if $a(x - x_0)$ is small, the forward error $x - x_0 \equiv {}^{1}/_{a} \cdot a(x - x_0)$ may be large given the ${}^{1}/_{a}$ factor.

We define the *condition number* to be a measure of a problem's sensitivity:

Definition 1.4 (Condition number)**.** The *condition number* of a problem is the ratio of how much its solution changes to the amount its statement changes under small perturbations. Alternatively, it is the ratio of forward to backward error for small changes in the problem statement.

Problems with small condition numbers are well-conditioned, and thus backward error can be used safely to judge success of approximate solution techniques. Contrastingly, much smaller backward error is needed to justify the quality of a candidate solution to a problem with a large condition number.

Example 1.8 ($ax = b$, part two)**.** Continuing Example 1.7, we can compute the condition number exactly:

$$c = \frac{\text{forward error}}{\text{backward error}} = \frac{x - x_0}{a(x - x_0)} \equiv \frac{1}{a}$$

In general, computing condition numbers is nearly as hard as computing forward error, and thus their exact computation is likely impossible. Even so, many times it is possible to find bounds or approximations for condition numbers to help evaluate how much a solution can be trusted.

Example 1.9 (Root-finding)**.** Suppose that we are given a smooth function $f : \mathbb{R} \to \mathbb{R}$ and want to find roots x with $f(x) = 0$. By Taylor's Theorem, $f(x + \varepsilon) \approx f(x) + \varepsilon f'(x)$ when $|\varepsilon|$ is small. Thus, an approximation of the condition number for finding the root x is given by

$$\frac{\text{change in forward error}}{\text{change in backward error}} = \frac{(x + \varepsilon) - x}{f(x + \varepsilon) - f(x)} \approx \frac{\varepsilon}{\varepsilon f'(x)} = \frac{1}{f'(x)}$$

This approximation generalizes the one in Example 1.8. If we do not know x we cannot evaluate $f'(x)$, but if we can examine the form of f and *bound* $|f'|$ near x, we have an idea of the worst-case situation.

Forward and backward error measure the *accuracy* of a solution. For the sake of scientific repeatability, we also wish to derive *stable* algorithms that produce self-consistent solutions to a class of problems. For instance, an algorithm that generates accurate solutions only one fifth of the time might not be worth implementing, even if we can use the techniques above to check whether a candidate solution is good. Other numerical methods require the client to tune several unintuitive parameters before they generate usable output and may be unstable or sensitive to changes to any of these options.

1.3 PRACTICAL ASPECTS

The theory of error analysis introduced in §1.2 will allow us to bound the quality of numerical techniques we introduce in future chapters. Before we proceed, it is worth noting some practical oversights and "gotchas" that pervade implementations of numerical methods.

We purposefully introduced the largest offender early in §1.1, which we repeat in a larger font for well-deserved emphasis:

Rarely if ever should the operator == and its equivalents be used on fractional values. Instead, some *tolerance* should be used to check if numbers are equal.

Finding a suitable replacement for `==` depends on particulars of the situation. Example 1.5 shows that a method for solving $A\vec{x} = \vec{b}$ can terminate when the residual $\vec{b} - A\vec{x}$ is zero; since we do not want to check if `A*x==b` explicitly, in practice implementations will check `norm(A*x-b)<epsilon`. This example demonstrates two techniques:

- The use of *backward* error $\vec{b} - A\vec{x}$ rather than forward error to determine when to terminate, and
- Checking whether backward error is less than `epsilon` to avoid the forbidden `==0` predicate.

The parameter `epsilon` depends on how accurate the desired solution must be as well as the quality of the discrete numerical system.

Based on our discussion of relative error, we can isolate another common cause of bugs in numerical software:

Beware of operations that transition between orders of magnitude, like division by small values and subtraction of similar quantities.

Catastrophic cancellation as in Example 1.4 can cause relative error to explode even if the inputs to an operation are known with near-complete certainty.

1.3.1 Computing Vector Norms

A programmer using floating-point data types and operations must be vigilant when it comes to detecting and preventing poor numerical operations. For example, consider the following code snippet for computing the norm $\|\vec{x}\|_2$ for a vector $\vec{x} \in \mathbb{R}^n$ represented as a 1D array `x[]`:

```
double normSquared = 0;
for (int i = 0; i < n; i++)
        normSquared += x[i]*x[i];
return sqrt(normSquared);
```

It is easy to see that in theory $\min_i |x_i| \leq \|\vec{x}\|_2/\sqrt{n} \leq \max_i |x_i|$, that is, the norm of $\vec{x}$ is on the order of the values of elements contained in $\vec{x}$. Hidden in the computation of $\|\vec{x}\|_2$, however, is the expression `x[i]*x[i]`. If there exists `i` such that `x[i]` is near `DOUBLE_MAX`, the product `x[i]*x[i]` will overflow even though $\|\vec{x}\|_2$ is still within the range of the `double`s. Such overflow is easily preventable by dividing $\vec{x}$ by its maximum value, computing the norm, and multiplying back:

```
function SIMPLE-SUM(x⃗)
    s ← 0                                   ▷ Current total
    for i ← 1, 2, ..., n : s ← s + x_i
    return s
```

(a)

```
function KAHAN-SUM(x⃗)
    s, c ← 0                                ▷ Current total and compensation
    for i ← 1, 2, ..., n
        v ← x_i + c                         ▷ Try to add x_i and compensation c to the sum
        s_next ← s + v                      ▷ Compute the summation result of this iteration
        ▷ Compute compensation using the Kahan error estimate
        c ← v − (s_next − s)
        s ← s_next                          ▷ Update sum
    return s
```

(b)

FIGURE 1.2 (a) A simplistic method for summing the elements of a vector $\vec{x}$; (b) the Kahan summation algorithm.

```
double maxElement = epsilon; // don't want to divide by zero!
for (int i = 0; i < n; i++)
        maxElement = max(maxElement, fabs(x[i]));
for (int i = 0; i < n; i++) {
        double scaled = x[i] / maxElement;
        normSquared += scaled*scaled;
}
return sqrt(normSquared) * maxElement;
```

The scaling factor alleviates the overflow problem by ensuring that elements being summed are no larger than 1, at the cost of additional computation time.

This small example shows one of many circumstances in which a single *character* of code can lead to a non-obvious numerical issue, in this case the product `*`. While our intuition from continuous mathematics is sufficient to formulate many numerical methods, we must always double-check that the operations we employ are valid when transitioning from theory to finite-precision arithmetic.

1.3.2 Larger-Scale Example: Summation

We now provide an example of a numerical issue caused by finite-precision arithmetic whose resolution involves a more subtle algorithmic trick. Suppose that we wish to sum a list of floating-point values stored in a vector $\vec{x} \in \mathbb{R}^n$, a task required by systems in accounting, machine learning, graphics, and nearly any other field. A simple strategy, iterating over the elements of $\vec{x}$ and incrementally adding each value, that appears in countless applications is detailed in Figure 1.2(a). For the vast majority of applications, this method is stable and mathematically valid, but in challenging cases it can fail.

What can go wrong? Consider the case where n is large and most of the values x_i are small and positive. Then, when i is large, the current sum s will be large relative to x_i. Eventually, s could be so large that adding x_i would change only the lowest-order bits of s,

and in the extreme case s could be large enough that adding x_i has no effect whatsoever. Put more simply, adding a long list of small numbers can result in a large sum, even if any single term of the sum could appear insignificant.

To understand this effect mathematically, suppose that computing a sum $a + b$ can be off by as much as a factor of $\varepsilon > 0$. Then, the method in Figure 1.2(a) can induce error on the order of $n\varepsilon$, which grows linearly with n. In fact, if most elements x_i are on the order of ε, then the sum cannot be trusted *whatsoever*! This is a disappointing result: The error can be as large as the sum itself.

Fortunately, there are many ways to do better. For example, adding the smallest values first might make sure they are not deemed insignificant. Methods recursively adding pairs of values from $\vec{x}$ and building up a sum also are more stable, but they can be difficult to implement as efficiently as the **for** loop above. Thankfully, an algorithm by Kahan provides an easily-implemented "compensated summation" method that is nearly as fast [64].

The useful observation to make is that we can approximate the inaccuracy of s as it changes from iteration to iteration. To do so, consider the expression

$$((a + b) - a) - b.$$

Algebraically, this expression equals zero. Numerically, however, this may not be the case. In particular, the sum $(a+b)$ may be rounded to floating-point precision. Subtracting a and b one-at-a-time then yields an approximation of the error of approximating $a+b$. Removing a and b from $a + b$ intuitively transitions *from* large orders of magnitude *to* smaller ones rather than vice versa and hence is less likely to induce truncation error than evaluating the sum $a + b$; this observation explains why the error estimate is not itself as prone to rounding issues as the original operation.

With this observation in mind, the Kahan technique proceeds as in Figure 1.2(b). In addition to maintaining the sum s, now we keep track of a *compensation* value c approximating the difference between s and the true sum at each iteration i. During each iteration, we attempt to add this compensation to s in addition to the current element x_i of $\vec{x}$; then we recompute c to account for the latest error.

Analyzing the Kahan algorithm requires more careful bookkeeping than analyzing the incremental technique in Figure 1.2(a). Although constructing a formal mathematical argument is outside the scope of our discussion, the final mathematical result is that error is on the order $O(\varepsilon + n\varepsilon^2)$, a considerable improvement over $O(n\varepsilon)$ when $0 < \varepsilon \ll 1$. Intuitively it makes sense that the $O(n\varepsilon)$ term from Figure 1.2(a) is reduced, since the compensation attempts to represent the small values that were otherwise neglected. Formal arguments for the ε^2 bound are surprisingly involved; one detailed derivation can be found in [45].

Implementing Kahan summation is straightforward but more than doubles the operation count of the resulting program. In this way, there is an implicit trade-off between speed and accuracy that software engineers must make when deciding which technique is most appropriate. More broadly, Kahan's algorithm is one of several methods that bypass the accumulation of numerical error during the course of a computation consisting of more than one operation. Another representative example from the field of computer graphics is Bresenham's algorithm for rasterizing lines [16], which uses only integer arithmetic to draw lines even when they intersect rows and columns of pixels at non-integer locations.

1.4 EXERCISES

1.1 When might it be preferable to use a fixed-point representation of real numbers over floating-point? When might it be preferable to use a floating-point representation of real numbers over fixed-point?

1.2[DH] ("Extraterrestrial chemistry") Suppose we are programming a planetary rover to analyze the chemicals in a gas found on a neighboring planet. Our rover is equipped with a flask of volume $0.5\,\mathrm{m}^3$ and also has pressure and temperature sensors. Using the sensor readouts from a given sample, we would like our rover to determine the amount of gas our flask contains.

One of the fundamental physical equations describing a gas is the Ideal Gas Law $PV = nRT$, which states:

$$(P)\text{ressure} \cdot (V)\text{olume} = \text{amou}(n)\text{t of gas} \cdot R \cdot (T)\text{emperature},$$

where R is the ideal gas constant, approximately equal to $8.31\,\mathrm{J} \cdot \mathrm{mol}^{-1} \cdot \mathrm{K}^{-1}$. Here, P is in pascals, V is in cubic meters, n is in moles, and T is in Kelvin. We will use this equation to approximate n given the other variables.

(a) Describe any forms of truncation, discretization, modeling, and input error that can occur when solving this problem.

(b) Our rover's pressure and temperature sensors do not have perfect accuracy. Assume the pressure and temperature sensor measurements are accurate to within $\pm\varepsilon_P$ and $\pm\varepsilon_T$, resp. Assuming V, R, and fundamental arithmetic operations like $+$ and $\times$ induce no errors, what is the forward error in computing n?

(c) Continuing the previous part, suppose $P = 100\,\mathrm{Pa}$, $T = 300\,\mathrm{K}$, $\varepsilon_P = 1\,\mathrm{Pa}$, and $\varepsilon_T = 0.5\,\mathrm{K}$. What are the worst absolute and relative errors that we could obtain from a computation of n (i.e., upper bound these errors)?

(d) Experiment with perturbing the variables P and T. Based on how much your estimate of n changes between the experiments, can you gain a sense of whether this problem is well-conditioned or ill-conditioned?

1.3[DH] In many numerical problems, we wish to evaluate function f at a point x, obtaining the output $y = f(x)$. Assuming f and its derivative are continuous and exist everywhere, what is the condition number for numerically computing $f(x)$? Give an example of an f that, at least in a certain range of x, yields a large condition number. Do the same for a small condition number.
Hint: Use the derivative "difference quotient" and Taylor's Theorem to obtain an answer to the first part in terms of x, $f(x)$, and $f'(x)$.

1.4 Suppose $f : \mathbb{R} \to \mathbb{R}$ is infinitely differentiable, and we wish to write algorithms for finding x^* minimizing $f(x)$. Our algorithm outputs x_{est}, an approximation of x^*. Assuming that in our context this problem is equivalent to finding roots of $f'(x)$, write expressions for:

(a) Forward error of the approximation.

(b) Backward error of the approximation.

(c) Conditioning of this minimization problem near x^*.

1.5 (adapted from Stanford CS 205A, 2012) Thanks to floating-point arithmetic, in most implementations of numerical algorithms we cannot expect that computations involving fractional values can be carried out with 100% precision. Instead, every time we do a numerical operation we induce the potential for error. Many models exist for

studying how this error affects the quality of a numerical operation; in this problem, we will explore one common model.

Suppose we care about an operation $\diamond$ between two scalars x and y; here $\diamond$ might stand for $+$, $-$, $\times$, $\div$, and so on. As a model for the error that occurs when computing $x \diamond y$, we will say that evaluating $x \diamond y$ on the computer yields a number $(1+\varepsilon)(x \diamond y)$ for some number ε satisfying $0 \leq |\varepsilon| < \varepsilon_{\max} \ll 1$; we will assume ε can depend on $\diamond$, x, and y.

(a) Why is this a reasonable model for modeling numerical issues in floating point arithmetic? For example, why does this make more sense than assuming that the output of evaluating $x \diamond y$ is $(x \diamond y) + \varepsilon$?

(b) Suppose we are given two vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$ and compute their dot product as s_n via the recurrence:

$$s_0 \equiv 0$$
$$s_k \equiv s_{k-1} + x_k y_k.$$

In practice, both the addition and multiplication steps of computing s_k from s_{k-1} induce numerical error. Use $\hat{s}_k$ to denote the actual value computed incorporating numerical error, and denote $e_k \equiv |\hat{s}_k - s_k|$. Show that

$$|e_n| \leq n\varepsilon_{\max} \sum_{k=1}^{n} |x_k||y_k| + O(n\varepsilon_{\max}^2).$$

1.6 Argue using the error model from the previous problem that the relative error of computing $x - y$ for $x, y > 0$ can be unbounded. This phenomenon is known as "catastrophic cancellation" and can cause serious numerical issues.

1.7 In this problem, we continue to explore the conditioning of root-finding. Suppose $f(x)$ and $p(x)$ are smooth functions of $x \in \mathbb{R}$.

(a) Thanks to inaccuracies in how we evaluate or express $f(x)$, we might accidentally compute roots of a perturbation $f(x) + \varepsilon p(x)$. Take x^* to be a root of f, so $f(x^*) = 0$. If $f'(x^*) \neq 0$, for small ε we can write a function $x(\varepsilon)$ such that $f(x(\varepsilon)) + \varepsilon p(x(\varepsilon)) = 0$, with $x(0) = x^*$. Assuming such a function exists and is differentiable, show:

$$\left.\frac{dx}{d\varepsilon}\right|_{\varepsilon=0} = -\frac{p(x^*)}{f'(x^*)}.$$

(b) Assume $f(x)$ is given by Wilkinson's polynomial [125]:

$$f(x) \equiv (x-1) \cdot (x-2) \cdot (x-3) \cdot \cdots \cdot (x-20).$$

We could have expanded $f(x)$ in the monomial basis as $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{20} x^{20}$, for appropriate choices of $a_0, \ldots, a_{20}$. If we express the coefficient a_{19} inaccurately, we could use the model from part 1.7a with $p(x) \equiv x^{19}$ to predict how much root-finding will suffer. For these choices of $f(x)$ and $p(x)$, show:

$$\left.\frac{dx}{d\varepsilon}\right|_{\varepsilon=0, x^*=j} = -\prod_{k \neq j} \frac{j}{j-k}.$$

(c) Compare $\frac{dx}{d\varepsilon}$ from the previous part for $x^* = 1$ and $x^* = 20$. Which root is more stable to this perturbation?

1.8 The roots of the quadratic function $ax^2 + bx + c$ are given by the *quadratic equation*

$$x^* \in \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

(a) Prove the alternative formula

$$x^* \in \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

(b) Suppose we are writing software to find these roots x^*. For which values of a, b, or c would the original or alternative formula be most stable numerically?

1.9 One technique for tracking uncertainty in a calculation is the use of *interval arithmetic*. In this system, an uncertain value for a variable x is represented as the interval $[x] \equiv [\underline{x}, \overline{x}]$ representing the range of possible values for x, from $\underline{x}$ to $\overline{x}$. Assuming infinite-precision arithmetic, give update rules for the following in terms of $\underline{x}$, $\overline{x}$, $\underline{y}$, and $\overline{y}$:

- $[x] + [y]$
- $[x] - [y]$
- $[x] \times [y]$
- $[x] \div [y]$
- $[x]^{1/2}$

Additionally, propose a conservative modification for finite-precision arithmetic.

1.10 Algorithms for dealing with geometric primitives such as line segments and triangles are notoriously difficult to implement in a numerically-stable fashion. Here, we highlight a few ideas from "ε-geometry," a technique built to deal with these issues [51].

(a) Take $\vec{p}, \vec{q}, \vec{r} \in \mathbb{R}^2$. Why might it be difficult to determine whether $\vec{p}$, $\vec{q}$, and $\vec{r}$ are collinear using finite-precision arithmetic?

(b) We will say $\vec{p}$, $\vec{q}$, and $\vec{r}$ are ε-collinear if there exist $\vec{p}'$ with $\|\vec{p} - \vec{p}'\|_2 \leq \varepsilon$, $\vec{q}'$ with $\|\vec{q} - \vec{q}'\|_2 \leq \varepsilon$, and $\vec{r}'$ with $\|\vec{r} - \vec{r}'\|_2 \leq \varepsilon$ such that $\vec{p}'$, $\vec{q}'$, and $\vec{r}'$ are exactly collinear. For fixed $\vec{p}$ and $\vec{q}$, sketch the region $\{\vec{r} \in \mathbb{R}^2 : \vec{p}, \vec{q}, \vec{r}$ are ε-collinear$\}$. This region is known as the *ε-butterfly* of $\vec{p}$ and $\vec{q}$.

(c) An ordered triplet $(\vec{p}, \vec{q}, \vec{r}) \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$ is ε-clockwise if the three points can be perturbed by at most distance ε so that they form a triangle whose vertices are in clockwise order; we will consider collinear triplets to be both clockwise and counterclockwise. For fixed $\vec{p}$ and $\vec{q}$, sketch the region $\{\vec{r} \in \mathbb{R}^2 : (\vec{p}, \vec{q}, \vec{r})$ is ε-clockwise$\}$.

(d) Show a triplet is ε-collinear if and only if it is both ε-clockwise and ε-counterclockwise.

(e) A point $\vec{x} \in \mathbb{R}^2$ is ε-inside the triangle $(\vec{p}, \vec{q}, \vec{r})$ if and only if $\vec{p}$, $\vec{q}$, $\vec{r}$, and $\vec{x}$ can be moved by at most distance ε such that the perturbed $\vec{x}'$ is exactly inside the perturbed triangle $(\vec{p}', \vec{q}', \vec{r}')$. Show that when $\vec{p}$, $\vec{q}$, and $\vec{r}$ are in (exactly) clockwise order, $\vec{x}$ is ε-inside $(\vec{p}, \vec{q}, \vec{r})$ if and only if $(\vec{p}, \vec{q}, \vec{x})$, $(\vec{q}, \vec{r}, \vec{x})$, and $(\vec{r}, \vec{p}, \vec{x})$ are all ε-clockwise.

II

Linear Algebra

CHAPTER 2

Linear Systems and the LU Decomposition

CONTENTS

2.1 Solvability of Linear Systems 43
2.2 Ad-Hoc Solution Strategies 45
2.3 Encoding Row Operations 47
2.3.1 Permutation 47
2.3.2 Row Scaling 48
2.3.3 Elimination 48
2.4 Gaussian Elimination 50
2.4.1 Forward Substitution 51
2.4.2 Back Substitution 52
2.4.3 Analysis of Gaussian Elimination 53
2.5 LU Factorization 55
2.5.1 Constructing the Factorization 56
2.5.2 Using the Factorization 56
2.5.3 Implementing LU 57

We commence our discussion of numerical algorithms by deriving ways to solve the linear system of equations $A\vec{x} = \vec{b}$. We will explore direct applications of these systems in Chapter 3, showing a variety of computational problems that can be approached by constructing appropriate A and $\vec{b}$ and solving for $\vec{x}$. Furthermore, linear solves will serve as a basic step in larger methods for optimization, simulation, and other numerical tasks considered in almost all future chapters. For these reasons, a thorough treatment and understanding of linear systems is critical.

2.1 SOLVABILITY OF LINEAR SYSTEMS

As introduced in §0.3.4, systems of linear equations like

$$\begin{aligned} 3x + 2y &= 6 \\ -4x + y &= 7 \end{aligned}$$

can be written in matrix form as in

$$\begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \end{pmatrix}.$$

More generally, we can write linear systems in the form $A\vec{x} = \vec{b}$ for $A \in \mathbb{R}^{m\times n}$, $\vec{x} \in \mathbb{R}^n$, and $\vec{b} \in \mathbb{R}^m$.

The solvability of $A\vec{x} = \vec{b}$ must fall into one of three cases:

1. The system may not admit any solutions, as in:

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

   This system enforces x two incompatible conditions simultaneously: $x = -1$ and $x = 1$.

2. The system may admit a single solution; for instance, the system at the beginning of this section is solved by $(x, y) = (-8/11, 45/11)$.

3. The system may admit infinitely many solutions, e.g. $0\vec{x} = \vec{0}$. If a system $A\vec{x} = \vec{b}$ admits two distinct solutions $\vec{x}_0$ and $\vec{x}_1$, then it automatically has infinitely many solutions of the form $c\vec{x}_0 + (1 - c)\vec{x}_1$ for $c \in \mathbb{R}$, since

$$A(c\vec{x}_0 + (1 - c)\vec{x}_1) = cA\vec{x}_0 + (1 - c)A\vec{x}_1 = c\vec{b} + (1 - c)\vec{b} = \vec{b}.$$

   Because it has multiple solutions, this linear system is labeled *underdetermined*.

The solvability of the system $A\vec{x} = \vec{b}$ depends both on A and on $\vec{b}$. For instance, if we modify the unsolvable system above to

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

then the system changes from having no solutions to infinitely many of the form $(1, y)$. Every matrix A admits a right hand side $\vec{b}$ such that $A\vec{x} = \vec{b}$ is solvable, since $A\vec{x} = \vec{0}$ always can be solved by $\vec{x} \equiv \vec{0}$ regardless of A.

For alternative intuition about the solvability of linear systems, recall from §0.3.1 that the matrix-vector product $A\vec{x}$ can be viewed as a linear combination of the columns of A with weights from $\vec{x}$. Thus, as mentioned in §0.3.4, we can expect $A\vec{x} = \vec{b}$ to be solvable exactly when $\vec{b}$ is in the column space of A.

In a broad way, the "shape" of the matrix $A \in \mathbb{R}^{m\times n}$ has considerable bearing on the solvability of $A\vec{x} = \vec{b}$. First, consider the case when A is "wide," that is, when it has more columns than rows ($n > m$). Each column is a vector in $\mathbb{R}^m$, so at most the column space can have dimension m. Since $n > m$, the n columns of A must then be linearly dependent; this implies that there exists a set of weights $\vec{x}_0 \neq \vec{0}$ such that $A\vec{x}_0 = \vec{0}$. Then, if we can solve $A\vec{x} = \vec{b}$ for $\vec{x}$, then $A(\vec{x} + \alpha\vec{x}_0) = A\vec{x} + \alpha A\vec{x}_0 = \vec{b} + \vec{0} = \vec{b}$, showing that there are actually infinitely many solutions $\vec{x}$ to $A\vec{x} = \vec{b}$. In other words:

No wide matrix system admits a unique solution.

When A is "tall," that is, when it has more rows than columns ($m > n$), then its n columns cannot possibly span the larger-dimensional $\mathbb{R}^m$. For this reason, there exists some vector $\vec{b}_0 \in \mathbb{R}^m \backslash \operatorname{col} A$. By definition, this $\vec{b}_0$ cannot satisfy $A\vec{x} = \vec{b}_0$ for *any* $\vec{x}$. That is:

For every tall matrix A, there exists a $\vec{b}_0$ such that $A\vec{x} = \vec{b}_0$ is not solvable.

The situations above are far from favorable for designing numerical algorithms. In the wide case, if a linear system admits many solutions we must specify *which* solution is desired by the user. After all, the solution $\vec{x} + 10^{31}\vec{x}_0$ might not be as meaningful as $\vec{x} - 0.1\vec{x}_0$. In the tall case, even if $A\vec{x} = \vec{b}$ is solvable for a particular $\vec{b}$, a small perturbation $A\vec{x} = \vec{b} + \varepsilon\vec{b}_0$ may not be solvable; even the rounding procedures discussed in the last chapter can move a tall system from solvable to unsolvable.

Given these complications, in this chapter we will make some simplifying assumptions:

- We will consider only *square* $A \in \mathbb{R}^{n\times n}$.
- We will assume that A is *nonsingular*, that is, that $A\vec{x} = \vec{b}$ is solvable for any $\vec{b}$.

From §0.3.4, the nonsingularity condition ensures that the columns of A span $\mathbb{R}^n$ and implies the existence of a matrix A^{-1} satisfying $A^{-1}A = AA^{-1} = I_{n\times n}$. We will relax these conditions in subsequent chapters.

A misleading observation is to think that solving $A\vec{x} = \vec{b}$ is equivalent to computing the matrix A^{-1} explicitly and then multiplying to find $\vec{x} \equiv A^{-1}\vec{b}$. While this formula is valid mathematically, it can represent a considerable amount of overkill and potential for numerical instability for several reasons:

- The matrix A^{-1} may contain values that are difficult to express in floating-point precision, in the same way that $1/\varepsilon \to \infty$ as $\varepsilon \to 0$.
- It may be possible to tune the solution strategy both to A and to $\vec{b}$, e.g. by working with the columns of A that are the closest to $\vec{b}$ first. Strategies like these can provide higher numerical stability.
- Even if A is sparse, meaning it contains many zero values that do not need to be stored explicitly, or has other special structure, the same may not be true for A^{-1}.

We highlight this point as a common source of error and inefficiency in numerical software:

Avoid computing A^{-1} explicitly unless you have a strong justification for doing so.

2.2 AD-HOC SOLUTION STRATEGIES

In introductory algebra, we often approach the problem of solving a linear system of equations as a puzzle rather than as a mechanical exercise. The strategy is to "isolate" variables, iteratively simplifying individual equalities until each is of the form $x = \text{const}$. To formulate step-by-step algorithms for solving linear systems, it is instructive to carry out an example of this methodology with an eye for aspects that can be fashioned into a general technique.

We will consider the following system:

$$\begin{aligned} y - z &= -1 \\ 3x - y + z &= 4 \\ x + y - 2z &= -3. \end{aligned}$$

Alongside each simplification step, we will maintain a matrix system encoding the current

state. Rather than writing out $A\vec{x} = \vec{b}$ explicitly, we can save a bit of space by writing the "augmented" matrix below:

$$\left(\begin{array}{ccc|c} 0 & 1 & -1 & -1 \\ 3 & -1 & 1 & 4 \\ 1 & 1 & -2 & -3 \end{array}\right)$$

We can write linear systems this way so long as we agree that variable coefficients remain on the left of the line and the constants on the right.

Perhaps we wish to deal with the variable x first. For convenience, we may *permute* the rows of the system so that the third equation appears first:

$$\begin{aligned} x + y - 2z &= -3 \\ y - z &= -1 \\ 3x - y + z &= 4 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 1 & -2 & -3 \\ 0 & 1 & -1 & -1 \\ 3 & -1 & 1 & 4 \end{array}\right)$$

We can then *substitute* the first equation into the third to eliminate the $3x$ term. This is the same as scaling the relationship $x + y - 2z = -3$ by -3 and adding the result to the third equation:

$$\begin{aligned} x + y - 2z &= -3 \\ y - z &= -1 \\ -4y + 7z &= 13 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 1 & -2 & -3 \\ 0 & 1 & -1 & -1 \\ 0 & -4 & 7 & 13 \end{array}\right)$$

Similarly, to eliminate y from the third equation we can multiply the second equation by 4 and add the result to the third:

$$\begin{aligned} x + y - 2z &= -3 \\ y - z &= -1 \\ 3z &= 9 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 1 & -2 & -3 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 3 & 9 \end{array}\right)$$

We have now isolated z! Thus, we can scale the third row by $1/3$ to yield an expression for z:

$$\begin{aligned} x + y - 2z &= -3 \\ y - z &= -1 \\ z &= 3 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 1 & -2 & -3 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

Now, we can substitute $z = 3$ into the other two equations to remove z from all but the final row:

$$\begin{aligned} x + y &= 3 \\ y &= 2 \\ z &= 3 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

Finally, we make a similar substitution for y to complete the solve:

$$\begin{aligned} x &= 1 \\ y &= 2 \\ z &= 3 \end{aligned} \qquad \left(\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right)$$

Revisiting the steps above yields a few observations about how to solve linear systems:

- We wrote successive systems $A_i\vec{x} = \vec{b}_i$ that can be viewed as simplifications of the original $A\vec{x} = \vec{b}$.
- We solved the system without ever writing down A^{-1}.

- We repeatedly used a few elementary operations: scaling, adding, and permuting rows.
- The same operations were applied to A and $\vec{b}$. If we scaled the k-th row of A, we also scaled the k-th row of $\vec{b}$. If we added rows k and ℓ of A, we added rows k and ℓ of $\vec{b}$.
- The steps did not depend on $\vec{b}$. That is, all of our decisions were motivated by eliminating nonzero values in A; $\vec{b}$ just came along for the ride.
- We terminated when we reached the simplified system $I_{n\times n}\vec{x} = \vec{b}$.

We will use all of these general observations about solving linear systems to our advantage.

2.3 ENCODING ROW OPERATIONS

Looking back at the example in §2.2, we see that solving $A\vec{x} = \vec{b}$ only involved three operations: permutation, row scaling, and adding the scale of one row to another. We can solve *any* linear system this way, so it is worth exploring these operations in more detail.

A pattern we will see for the remainder of this chapter is the use of matrices to express row operations. For example, the following two descriptions of an operation on a matrix A are equivalent:

1. Scale the first row of A by 2
2. Replace A with S_2A, where S_2 is defined by:

$$S_2 \equiv \begin{pmatrix} 2 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

When presenting the theory of matrix simplification, it is cumbersome to use words to describe each operation as in the first option, so when possible we will encode matrix algorithms as a series of pre- and post-multiplications by specially-designed matrices like S_2 above.

This description in terms of matrices, however, is a *theoretical* construction. Implementations of algorithms for solving linear systems absolutely should not construct matrices like S_2 explicitly and then use generic matrix multiplication machinery to apply them. For example, if $A \in \mathbb{R}^{n\times n}$, it should take n steps to scale the first row of A by 2, but explicitly constructing $S_2 \in \mathbb{R}^{n\times n}$ and applying it to A takes n^3 steps! That is, we will show for notational convenience that row operations *can* be encoded using matrix multiplication, but do not *have* to be.

2.3.1 Permutation

Our first step in §2.2 was to swap two of the rows. More generally, we might index the rows of a matrices using the numbers $1, \ldots, m$. Then, a *permutation* of those rows can be written as a function $\sigma : \{1, \ldots, m\} \to \{1, \ldots, m\}$ such that $\{\sigma(1), \ldots, \sigma(m)\} = \{1, \ldots, m\}$, that is, σ maps every index to a different target.

If $\vec{e}_k$ is the k-th standard basis vector, the product $\vec{e}_k^\top A$ is the k-th row of the matrix

A. We can "stack" or concatenate these row vectors vertically to yield a matrix permuting the rows according to σ:

$$P_\sigma \equiv \begin{pmatrix} - & \vec{e}_{\sigma(1)}^\top & - \\ - & \vec{e}_{\sigma(2)}^\top & - \\ & \vdots & \\ - & \vec{e}_{\sigma(m)}^\top & - \end{pmatrix}$$

The product $P_\sigma A$ is the matrix A with rows permuted according to σ.

Example 2.1 (Permutation matrices)**.** Suppose we wish to permute rows of a matrix in $\mathbb{R}^{3\times 3}$ with $\sigma(1) = 2$, $\sigma(2) = 3$, and $\sigma(3) = 1$. According to our formula we have

$$P_\sigma = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

From Example 2.1, we can see that P_σ has ones in positions indexed $(k, \sigma(k))$ and zeros elsewhere. Reversing the order of each pair, that is, putting ones in positions indexed $(\sigma(k), k)$ and zeros elsewhere, undoes the effect of the permutation. Hence, the inverse of P_σ must be its transpose $P_\sigma^\top$. Symbolically, we can write $P_\sigma^\top P_\sigma = I_{m\times m}$, or equivalently $P_\sigma^{-1} = P_\sigma^\top$.

2.3.2 Row Scaling

Suppose we write down a list of constants $a_1, \ldots, a_m$ and seek to scale the k-th row of A by a_k for each k. This task is accomplished by applying the scaling matrix S_a:

$$S_a \equiv \begin{pmatrix} a_1 & 0 & 0 & \cdots \\ 0 & a_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_m \end{pmatrix}$$

Assuming that all a_k satisfy $a_k \neq 0$, it is easy to invert S_a by scaling back:

$$S_a^{-1} = S_{1/a} \equiv \begin{pmatrix} 1/a_1 & 0 & 0 & \cdots \\ 0 & 1/a_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/a_m \end{pmatrix}$$

If any a_k is zero, S_a is not invertible.

2.3.3 Elimination

Finally, suppose we wish to scale row k by a constant c and add the result to row ℓ; we will assume $k \neq \ell$. This operation may seem less natural than the previous two, but actually it is quite practical. In particular, it is the only one we need to combine equations from different rows of the linear system! We again will realize this operation using an "elimination matrix" M such that the product MA is the result of applying this operation to matrix A.

The product $\vec{e}_k^\top A$ picks out the k-th row of A. Premultiplying the result by $\vec{e}_\ell$ yields a matrix $\vec{e}_\ell \vec{e}_k^\top A$ that is zero except on its ℓ-th row, which is equal to the k-th row of A.

Example 2.2 (Elimination matrix construction)**.** Take

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Suppose we wish to isolate the third row of $A \in \mathbb{R}^{3\times 3}$ and move it to row two. As discussed above, this operation is accomplished by writing:

$$\begin{aligned} \vec{e}_2\vec{e}_3^{\top} A &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 7 & 8 & 9 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 7 & 8 & 9 \\ 0 & 0 & 0 \end{pmatrix} \end{aligned}$$

We multiplied right-to-left above but just as easily could have grouped the product as $(\vec{e}_2\vec{e}_3^{\top})A$. The structure of this product is:

$$\vec{e}_2\vec{e}_3^{\top} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

We have succeeded in isolating row k and moving it to row ℓ. Our original elimination operation was to add c times row k to row ℓ, which we can now carry out using the sum $A + c\vec{e}_\ell\vec{e}_k^{\top} A = (I_{n\times n} + c\vec{e}_\ell\vec{e}_k^{\top})A$. In summary, the desired elimination matrix is $M \equiv I_{n\times n} + c\vec{e}_\ell\vec{e}_k^{\top}$.

The action of M is easy to undo: Scale row k by c and *subtract* the result from row ℓ. We can check this formally:

$$\begin{aligned} (I_{n\times n} + c\vec{e}_\ell\vec{e}_k^{\top})(I_{n\times n} - c\vec{e}_\ell\vec{e}_k^{\top}) &= I_{n\times n} + (-c\vec{e}_\ell\vec{e}_k^{\top} + c\vec{e}_\ell\vec{e}_k^{\top}) - c^2\vec{e}_\ell\vec{e}_k^{\top}\vec{e}_\ell\vec{e}_k^{\top} \\ &= I_{n\times n} - c^2\vec{e}_\ell(\vec{e}_k^{\top}\vec{e}_\ell)\vec{e}_k^{\top} \\ &= I_{n\times n} \text{ since } \vec{e}_k^{\top}\vec{e}_\ell = \vec{e}_k \cdot \vec{e}_\ell, \text{ and } k \neq \ell \end{aligned}$$

That is, $M^{-1} = I_{n\times n} - c\vec{e}_\ell\vec{e}_k^{\top}$.

Example 2.3 (Solving a system)**.** We can now encode each of our operations from Section 2.2 using the matrices we have constructed above:

1. Permute the rows to move the third equation to the first row:

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

2. Scale row one by -3 and add the result to row three:

$$E_1 = I_{3\times 3} - 3\vec{e}_3\vec{e}_1^{\top} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}$$

3. Scale row two by 4 and add the result to row three:

$$E_2 = I_{3\times 3} + 4\vec{e}_3\vec{e}_2^{\top} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix}$$

4. Scale row three by $1/3$:

$$S = \mathrm{diag}(1, 1, 1/3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}$$

5. Scale row three by 2 and add it to row one:

$$E_3 = I_{3\times 3} + 2\vec{e}_1\vec{e}_3^{\top} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

6. Add row three to row two:

$$E_4 = I_{3\times 3} + \vec{e}_2\vec{e}_3^{\top} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

7. Scale row three by -1 and add the result to row one:

$$E_5 = I_{3\times 3} - \vec{e}_1\vec{e}_3^{\top} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Thus, the inverse of A in Section 2.2 satisfies

$$\begin{aligned} A^{-1} &= E_5E_4E_3SE_2E_1P \\ &= \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/3 \end{pmatrix} \\ &\quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 4/3 & 1/3 & 0 \\ 7/3 & 1/3 & -1 \\ 4/3 & 1/3 & -1 \end{pmatrix}. \end{aligned}$$

Make sure you understand why these matrices appear in *reverse* order! As a reminder, we would not normally construct A^{-1} by multiplying the matrices above, since these operations can be implemented more efficiently than generic matrix multiplication. Even so, it is valuable to check that the theoretical operations we have defined are equivalent to the ones we have written in words.

2.4 GAUSSIAN ELIMINATION

The sequence of steps chosen in Section 2.2 was by no means unique: There are many different paths that can lead to the solution of $A\vec{x} = \vec{b}$. Our steps, however, used *Gaussian elimination*, a famous algorithm for solving linear systems of equations.

To introduce this algorithm, let's say our system has the following generic "shape:"

$$\left(\begin{array}{c|c} A & \vec{b} \end{array} \right) = \left(\begin{array}{cccc|c} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

Here, an $\times$ denotes a potentially nonzero value. The algorithm proceeds in phases described below.

2.4.1 Forward Substitution

Consider the upper-left element of our matrix:

$$\left(\begin{array}{c|c} A & \vec{b} \end{array} \right) = \left(\begin{array}{cccc|c} \textcircled{\times} & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

We will call this element the first *pivot* and will assume it is nonzero; if it is zero we can permute rows so that this is not the case. We first scale the first row by the reciprocal of the pivot so that the value in the pivot position is one:

$$\left(\begin{array}{cccc|c} \textcircled{1} & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{array} \right)$$

Now, we use the row containing the pivot to eliminate all other values underneath in the same column using the strategy in §2.3.3:

$$\left(\begin{array}{cccc|c} \textcircled{1} & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{array} \right)$$

At this point, the entire first column is zero below the pivot. We change the pivot label to the element in position $(2, 2)$ and repeat a similar series of operations to rescale the pivot row and use it to cancel the values underneath:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & \times & \times \\ 0 & \textcircled{1} & \times & \times & \times \\ 0 & 0 & \times & \times & \times \\ 0 & 0 & \times & \times & \times \end{array} \right)$$

At this point our matrix begins to gain some structure. After the first pivot has been eliminated from all other rows, the first column is zero except for the leading one. Thus,

function FORWARD-SUBSTITUTION($A, \vec{b}$)
▷ Converts a system $A\vec{x} = \vec{b}$ to an upper-triangular system $U\vec{x} = \vec{y}$.
▷ Assumes invertible $A \in \mathbb{R}^{n\times n}$ and $\vec{b} \in \mathbb{R}^n$.

$U, \vec{y} \leftarrow A, \vec{b}$ ▷ U will be upper triangular at completion
for $p \leftarrow 1, 2, \ldots, n$ ▷ Iterate over current pivot row p
  $q \leftarrow$ FIND-PIVOT(U,p) ▷ Finds $q \geq p$ such that $u_{qp} \neq 0$
  SWAP(y_p,y_q) ▷ Swap rows p and q
  for $c \leftarrow p, \ldots, n$:SWAP(u_{pc},u_{qc})
  $s \leftarrow 1/u_{pp}$ ▷ Scale row to make pivot equal one
  $y_p \leftarrow s \cdot y_p$
  for $c \leftarrow p, \ldots, n : u_{pc} \leftarrow s \cdot u_{pc}$
  for $r \leftarrow p, \ldots, n$ ▷ Eliminate from future rows
    $s \leftarrow -u_{rp}$ ▷ Scale row p by s and add to row r
    $y_r \leftarrow y_r + s \cdot y_p$
    for $c \leftarrow p, \ldots, n : u_{rc} \leftarrow u_{rc} + s \cdot u_{pc}$
return $U, \vec{y}$

FIGURE 2.1 Forward substitution for making matrix systems upper-triangular. More refined alternatives to FIND-PIVOT are discussed in §2.4.3.

any row operation involving rows two to m will not affect the zeros in column one. Similarly, after the second pivot has been processed, operations on rows three to m will not remove the zeros in columns one and two.

We repeat this process until the matrix becomes *upper-triangular*:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & \times & \times \\ 0 & 1 & \times & \times & \times \\ 0 & 0 & 1 & \times & \times \\ 0 & 0 & 0 & \textcircled{1} & \times \end{array}\right)$$

The method above of making a matrix upper triangular is known as *forward substitution* and is detailed in Figure 2.1.

2.4.2 Back Substitution

Eliminating the remaining $\times$'s from the remaining upper-triangular system is an equally straightforward process proceeding in *reverse* order of rows and eliminating backward. After the first set of back substitution steps, we are left with the following shape:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & 0 & \times \\ 0 & 1 & \times & 0 & \times \\ 0 & 0 & 1 & 0 & \times \\ 0 & 0 & 0 & \textcircled{1} & \times \end{array}\right)$$

```
function BACK-SUBSTITUTION(U, y⃗)
  ▷ Solves upper-triangular systems U x⃗ = y⃗ for x⃗.
  x⃗ ← y⃗                          ▷ We will start from U x⃗ = y⃗ and simplify to I_{n×n} x⃗ = x⃗
  for p ← n, n − 1, ..., 1                          ▷ Iterate backward over pivots
    for r ← 1, 2, ..., p − 1                        ▷ Eliminate values above u_pp
      x_r ← x_r − u_rp x_p / u_pp
  return x⃗
```

FIGURE 2.2 Back-substitution for solving upper-triangular systems; this implementation returns the solution $\vec{x}$ to the system without modifying U.

Similarly, the second iteration yields:

$$\left(\begin{array}{cccc|c} 1 & \times & 0 & 0 & \times \\ 0 & 1 & 0 & 0 & \times \\ 0 & 0 & \textcircled{1} & 0 & \times \\ 0 & 0 & 0 & 1 & \times \end{array}\right)$$

After our final elimination step, we are left with our desired form:

$$\left(\begin{array}{cccc|c} \textcircled{1} & 0 & 0 & 0 & \times \\ 0 & 1 & 0 & 0 & \times \\ 0 & 0 & 1 & 0 & \times \\ 0 & 0 & 0 & 1 & \times \end{array}\right)$$

The right hand side now is the solution to the linear system $A\vec{x} = \vec{b}$. Figure 2.2 implements this method of *back-substitution* in more detail.

2.4.3 Analysis of Gaussian Elimination

Each row operation in Gaussian elimination—scaling, elimination, and swapping two rows—takes $O(n)$ time to complete, since you have to iterate over all n elements of a row (or two) of A. Once we choose a pivot, we have to do n forward- or back- substitutions into the rows below or above that pivot, resp.; this means the work for a single pivot in total is $O(n^2)$. In total, we choose one pivot per row, adding a final factor of n. Combining these counts, Gaussian elimination runs in $O(n^3)$ time.

One decision that takes place during Gaussian elimination that merits more discussion is the choice of *pivots*. We can permute rows of the linear system as we see fit before performing forward-substitution. This operation is *necessary* to be able to deal with all possible matrices A. For example, consider what would happen if we did not use pivoting on the following matrix:

$$A = \left(\begin{array}{cc} \textcircled{0} & 1 \\ 1 & 0 \end{array}\right)$$

The circled element is exactly zero, so we cannot expect to scale row one by any value to replace that 0 with a 1. This does *not* mean the system is not solvable—although singular matrices are guaranteed to have this issue—but rather it means we must *pivot* by swapping the first and second rows.

More generally, suppose A looks like:

$$A = \begin{pmatrix} \textcircled{\varepsilon} & 1 \\ 1 & 0 \end{pmatrix},$$

where $0 < \varepsilon \ll 1$. If we do not pivot, then the first iteration of Gaussian elimination yields:

$$\tilde{A} = \begin{pmatrix} \textcircled{1} & 1/\varepsilon \\ 0 & -1/\varepsilon \end{pmatrix},$$

We have transformed a matrix A that looks nearly like a permutation matrix ($A^{-1} \approx A^\top$, a very easy way to solve the system!) into a system with potentially **huge** values $1/\varepsilon$. This example is one of many instances in which we should try to avoid dividing by vanishingly small numbers.

In this way, there are cases when we may wish to pivot even when Gaussian elimination theoretically could proceed without such a step. Since we are scaling by the reciprocal of the pivot value, the most numerically stable option is to have a *large* pivot: Small pivots have large reciprocals, scaling matrix elements to large values in regimes that are likely to lose precision. There are two well-known pivoting strategies:

1. *Partial* pivoting looks through the current column and permutes rows of the matrix so that the largest absolute value appears on the diagonal.
2. *Full* pivoting iterates over the **entire** matrix and permutes both rows and columns to get the largest possible value on the diagonal. Permuting columns of a matrix is a valid operation: it corresponds to changing the labeling of the variables in the system, or post-multiplying A by a permutation.

Full pivoting is more expensive than partial pivoting since it requires iterating over the entire matrix (or using an advanced "priority queue"-type data structure) to find the largest absolute value, but it results in more numerical stability.

Example 2.4 (Pivoting)**.** Suppose after the first iteration of Gaussian elimination we are left with the following matrix:

$$\begin{pmatrix} 1 & 10 & -10 \\ 0 & \textcircled{0.1} & 9 \\ 0 & 4 & 6.2 \end{pmatrix}$$

If we implement partial pivoting, then we will look only in the second column and will swap the second and third rows; we leave the 10 in the first row since that row already has been visited during forward substitution:

$$\begin{pmatrix} 1 & 10 & -10 \\ 0 & \textcircled{4} & 6.2 \\ 0 & 0.1 & 9 \end{pmatrix}$$

If we implement full pivoting, then we will move the 9:

$$\begin{pmatrix} 1 & -10 & 10 \\ 0 & \textcircled{9} & 0.1 \\ 0 & 6.2 & 4 \end{pmatrix}$$

2.5 LU FACTORIZATION

There are many times when we wish to solve a sequence of problems $A\vec{x}_1 = \vec{b}_1, A\vec{x}_2 = \vec{b}_2, \ldots,$ where in each iteration the matrix A is the same. For example, in image processing we may want to apply the same filter encoded in A to a set of images encoded as $\vec{b}_1, \vec{b}_2, \ldots$. As we already have discussed, the steps of Gaussian elimination for solving $A\vec{x}_k = \vec{b}_k$ depend mainly on the structure of A rather than the values in a particular $\vec{b}_k$. Since A is kept constant here, we may wish to "remember" the steps we took to solve the system so that each time we are presented with a new $\vec{b}_k$ we do not have to start from scratch. This memory will be a compromise between restarting Gaussian elimination for each $\vec{b}_i$ and computing the potentially numerically-unstable inverse matrix A^{-1}.

Solidifying this suspicion that we can move some of the $O(n^3)$ for Gaussian elimination into precomputation time if we wish to reuse A, recall the *upper-triangular* system appearing after the forward substitution stage:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & \times & \times \\ 0 & 1 & \times & \times & \times \\ 0 & 0 & 1 & \times & \times \\ 0 & 0 & 0 & 1 & \times \end{array}\right)$$

Unlike forward substitution, solving this system by back-substitution only takes $O(n^2)$ time! Why? As implemented in Figure 2.2, back substitution can take advantage of the structure of the zeros in the system. For example, consider the circled elements of the initial upper-triangular system:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & \times & \times \\ 0 & 1 & \times & \times & \times \\ 0 & 0 & 1 & \times & \times \\ \textcircled{0} & \textcircled{0} & \textcircled{0} & 1 & \times \end{array}\right)$$

Since we know that the (circled) values to the left of the pivot are zero by definition of an upper-triangular matrix, we do not need to scale them or copy them upward explicitly. If we ignore these zeros completely, this step of backward substitution only takes n operations rather than the n^2 taken by the corresponding step of forward substitution.

Now, our next pivot benefits from a similar construction:

$$\left(\begin{array}{cccc|c} 1 & \times & \times & 0 & \times \\ 0 & 1 & \times & 0 & \times \\ \textcircled{0} & \textcircled{0} & 1 & \textcircled{0} & \times \\ 0 & 0 & 0 & 1 & \times \end{array}\right)$$

Again, the zeros on both sides of the one do not need to be copied explicitly.

A near-identical method can be used to solve *lower*-triangular systems of equations via forward substitution. Combining these observations, we have shown:

While Gaussian elimination takes $O(n^3)$ time, solving triangular systems takes $O(n^2)$ time.

In this section, we will revisit the steps of Gaussian elimination to show that they can be used to factorize the matrix A as $A = LU$, where L is lower-triangular and U is upper-triangular, so long as pivoting is not needed to solve $A\vec{x} = \vec{b}$. Once the matrices L and U are obtained, solving $A\vec{x} = \vec{b}$ can be carried out by instead solving $LU\vec{x} = \vec{b}$ using forward substitution followed by backward substitution; these two steps combined take $O(n^2)$ time rather than $O(n^3)$ time needed for full Gaussian elimination. This factorization also can be extended to a related and equally useful decomposition when pivoting is desired or necessary.

2.5.1 Constructing the Factorization

Other than full pivoting, from §2.3 we know that all the operations in Gaussian elimination can be thought of as pre-multiplying $A\vec{x} = \vec{b}$ by different matrices M to obtain easier systems $(MA)\vec{x} = M\vec{b}$. As demonstrated in Example 2.3, from this standpoint, each step of Gaussian elimination brings a new system $(M_k \cdots M_2 M_1 A)\vec{x} = M_k \cdots M_2 M_1 \vec{b}$. Explicitly storing these matrices M_k as $n \times n$ objects is overkill, but keeping this interpretation in mind from a theoretical perspective simplifies many of our calculations.

After the forward substitution phase of Gaussian elimination, we are left with an *upper triangular* matrix, which we can call $U \in \mathbb{R}^{n\times n}$. From the matrix multiplication perspective, we can write:

$$M_k \cdots M_1 A = U,$$

or, equivalently,

$$\begin{aligned} A &= (M_k \cdots M_1)^{-1} U \\ &= (M_1^{-1} M_2^{-1} \cdots M_k^{-1}) U \text{ from the fact } (AB)^{-1} = B^{-1}A^{-1} \\ &\equiv LU, \text{ if we make the definition } L \equiv M_1^{-1} M_2^{-1} \cdots M_k^{-1}. \end{aligned}$$

We know U is upper-triangular by design, but we have not characterized the structure of L; hence, our remaining task is to show that L is lower-triangular. To do so, recall that in the absence of pivoting, each matrix M_i is either a scaling matrix or has the structure $M_i = I_{n\times n} + c\vec{e}_\ell \vec{e}_k^\top$, from §2.3.3, where $\ell > k$ since we carried out forward substitution to obtain U. So, L is the product of scaling matrices and matrices of the form $M_i^{-1} = I_{n\times n} - c\vec{e}_\ell \vec{e}_k^\top$; these matrices are lower triangular since $\ell > k$. Since scaling matrices are diagonal, we know L is lower-triangular by the following proposition:

Proposition 2.1. The product of two or more upper-triangular matrices is upper-triangular, and the product of two or more lower-triangular matrices is lower-triangular.

Proof. Suppose A and B are upper triangular, and define $C \equiv AB$. By definition of upper triangular matrices, $a_{ij} = 0$ and $b_{ij} = 0$ when $i > j$. Now, fix two values i and j with $i > j$. Then,

$$\begin{aligned} c_{ij} &= \sum_k a_{ik} b_{kj} \text{ by definition of matrix multiplication} \\ &= a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj} \end{aligned}$$

The first $i - 1$ terms of the sum are zero because A is upper triangular, and the last $n - j$ terms are zero because B is upper triangular. Since $i > j$, $(i - 1) + (n - j) > n - 1$ and hence all n terms of the sum over k are zero, as needed.

If A and B are lower triangular, then $A^\top$ and $B^\top$ are upper triangular. By our proof above, $B^\top A^\top = (AB)^\top$ is upper triangular, showing that AB is again lower triangular. □

2.5.2 Using the Factorization

Now that we have factored $A = LU$, we can solve $A\vec{x} = \vec{b}$ in two steps, by writing $(LU)\vec{x} = \vec{b}$, or $\vec{x} = U^{-1}L^{-1}\vec{b}$:

1. Solve $L\vec{y} = \vec{b}$ for $\vec{y}$, yielding $\vec{y} = L^{-1}\vec{b}$.

2. With $\vec{y}$ now fixed, solve $U\vec{x} = \vec{y}$ for $\vec{x}$.

Checking the validity of $\vec{x}$ as a solution of the system $A\vec{x} = \vec{b}$ comes from the following chain of equalities:

$$\begin{aligned}\vec{x} &= U^{-1}\vec{y} \text{ from the second step} \\ &= U^{-1}(L^{-1}\vec{b}) \text{ from the first step} \\ &= (LU)^{-1}\vec{b} \text{ since } (AB)^{-1} = B^{-1}A^{-1} \\ &= A^{-1}\vec{b} \text{ since we factored } A = LU.\end{aligned}$$

The forward- and back-substitution steps to carry out the two steps above each take $O(n^2)$ time. So, given the LU factorization of A, solving $A\vec{x} = \vec{b}$ can be carried out faster than full $O(n^3)$ Gaussian elimination. When pivoting is necessary, we will modify our factorization to include a permutation matrix P to account for the swapped rows and/or columns, e.g. $A = LUP$; this minor change does not affect the asymptotic speed benefits of factorization, since $P^{-1} = P^\top$.

2.5.3 Implementing LU

The implementation of Gaussian elimination suggested in Figures 2.1 and 2.2 constructs U but not L. We can make some adjustments to factor $A = LU$ rather than solving a single system $A\vec{x} = \vec{b}$.

Let's examine what happens when we multiply two elimination matrices:

$$(I_{n\times n} - c_\ell \vec{e}_\ell \vec{e}_k^\top)(I_{n\times n} - c_p \vec{e}_p \vec{e}_k^\top) = I_{n\times n} - c_\ell \vec{e}_\ell \vec{e}_k^\top - c_p \vec{e}_p \vec{e}_k^\top$$

As in our construction of the inverse of an elimination matrix in §2.5.1, the remaining term vanishes since the standard basis is orthogonal. This formula shows that the product of elimination matrices used to forward-substitute a pivot after it is scaled to 1 has the form:

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \textcircled{1} & 0 & 0 \\ 0 & \times & 1 & 0 \\ 0 & \times & 0 & 1 \end{pmatrix},$$

where the values $\times$ are those used for forward substitutions of the circled pivot. Products of matrices of this form performed in forward-substitution order combine the values below the diagonal, as demonstrated in the following example:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 4 & 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 5 & 1 & 0 \\ 0 & 6 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 7 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 5 & 1 & 0 \\ 4 & 6 & 7 & 1 \end{pmatrix}$$

We construct U by pre-multiplying A by a sequence of elimination and scaling matrices. We can construct L simultaneously by *post*-multiplying by the inverses of these matrices, which can be accomplished easily by the observations above.

For any invertible diagonal matrix D, $(LD)(D^{-1}U)$ provides an alternative factorization of $A = LU$ into lower- and upper-triangular matrices. Thus, by rescaling we can decide to keep the elements along the diagonal of L in the LU factorization equal to 1. With this

```
function LU-Factorization-Compact(A)
  ▷ Factors A ∈ ℝ^{n×n} to A = LU in compact format.
  for p ← 1, 2, ..., n                    ▷ Choose pivots like in forward substitution
    for r ← p + 1, ..., n                 ▷ Forward substitution row
      s ← −a_rp / a_pp                    ▷ Amount to scale row p for forward substitution
      a_rp ← −s      ▷ L contains −s because it reverses the forward substitution
      for c ← p + 1, ..., n               ▷ Perform forward substitution
        a_rc ← a_rc + s a_pc
  return A
```

FIGURE 2.3 Pseudocode for computing the LU factorization of $A \in \mathbb{R}^{n \times n}$, stored in the compact $n \times n$ format described in §2.5.3. This algorithm will fail if pivoting is needed.

decision in place, we can compress our storage of *both* L and U into a single $n \times n$ matrix whose upper triangle is U and which is equal to L beneath the diagonal; the missing diagonal elements of L are all 1.

We are now ready to write pseudocode for LU factorization without pivoting, illustrated in Figure 2.3. This method extends the algorithm for forward substitution by storing the corresponding elements of L under the diagonal rather than zeros. This method has three nested loops and thus runs in $O(n^3) \approx \frac{2}{3}n^3$ time; after precomputing this factorization, however, solving $A\vec{x} = \vec{b}$ only takes $O(n^2)$ time using forward- and backward-substitution.

2.6 EXERCISES

2.1 Can *all* matrices $A \in \mathbb{R}^{n \times n}$ be factored $A = LU$? Why or why not?

2.2[DH] Factor the following matrix A as a product $A = LU$:

$$\begin{pmatrix} 1 & 2 & 7 \\ 3 & 5 & -1 \\ 6 & 1 & 4 \end{pmatrix}$$

2.3 Numerical algorithms appear in many components of simulation software for quantum physics. The Schrödinger equation and others involve *complex* numbers in $\mathbb{C}$, however, so we must extend the machinery we have developed for solving linear systems of equations to this case. Recall that a complex number $x \in \mathbb{C}$ can be written as $x = a + bi$, where $a, b \in \mathbb{R}$ and $i = \sqrt{-1}$. Suppose we wish to solve $A\vec{x} = \vec{b}$, but now $A \in \mathbb{C}^{n \times n}$ and $\vec{x}, \vec{b} \in \mathbb{C}^n$. Explain how a linear solver that takes only *real-valued* systems can be used to solve this equation.
Hint: Write $A = A_1 + A_2 i$, where $A_1, A_2 \in \mathbb{R}^{n \times n}$. Similarly decompose $\vec{x}$ and $\vec{b}$. In the end you will solve a $2n \times 2n$ real-valued system.

2.4 Show that if L is an invertible lower-triangular matrix, none of its diagonal elements can be zero. How does this lemma affect the construction in §2.5.3?

2.5 Show that the inverse of an (invertible) lower triangular matrix is lower triangular.

2.6 Show that any invertible matrix $A \in \mathbb{R}^{n\times n}$ with $a_{11} = 0$ cannot have a factorization $A = LU$ for lower triangular L and upper triangular U.

2.7 Show how the LU factorization of $A \in \mathbb{R}^{n\times n}$ can be used to compute the determinant of A.

2.8 For numerical stability and generality, we incorporated pivoting into our methods for Gaussian elimination. We can modify our construction of the LU factorization somewhat to incorporate pivoting as well.

(a) Argue that following the steps of Gaussian elimination on a matrix $A \in \mathbb{R}^{n\times n}$ with partial pivoting can be used to write $U = L_{n-1}P_{n-1}\cdots L_2P_2L_1P_1A$, where the P_i's are permutation matrices, the L_i's are lower-triangular, and U is upper-triangular.

(b) Show that P_i is a permutation matrix that swaps rows i and j for some $j > i$. Also, argue that L_i is the product of matrices of the form $I_{n\times n} + c\vec{e}_k\vec{e}_i^{\top}$ where $k > i$.

(c) Suppose $j, k > i$. Show $P_{jk}(I_{n\times n} + c\vec{e}_k\vec{e}_i^{\top}) = (I_{n\times n} + c\vec{e}_j\vec{e}_i^{\top})P_{jk}$, where P_{jk} is a permutation matrix swapping rows j and k.

(d) Combine the previous two parts to show that

$$L_{n-1}P_{n-1}\cdots L_2P_2L_1P_1 = L_{n-1}L'_{n-2}L'_{n-3}\cdots L'_1P_{n-1}\cdots P_2P_1,$$

where $L'_1, \ldots, L'_{n-2}$ are lower-triangular.

(e) Conclude that $A = PLU$, where P is a permutation matrix, L is lower-triangular, and U is upper-triangular.

(f) Extend the method from §2.5.2 for solving $A\vec{x} = \vec{b}$ when we have factored $A = PLU$, without affecting the time complexity compared to factorizations $A = LU$.

2.9 ("Block LU decomposition") Suppose a square matrix $M \in \mathbb{R}^{n\times n}$ is written in block form as

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where $A \in \mathbb{R}^{k\times k}$ is square and invertible.

(a) Show that we can decompose M as the product

$$M = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix}\begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix}\begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

Here I denotes an identity matrix of appropriate size.

(b) Suppose we decompose $A = L_1U_1$ and $D - CA^{-}1B = L_2U_2$. Show how to construct an LU factorization of M given these additional matrices.

(c) Use this structure to define a recursive algorithm for LU factorization; you can assume $n = 2^{\ell}$ for some $\ell > 0$. How does the efficiency of your method compare with the LU algorithm introduced in this chapter?

2.10 Suppose $A \in \mathbb{R}^{n\times n}$ is invertible and admits a factorization $A = LU$ with ones along the diagonal of L. Show that such a decomposition of A is unique.

CHAPTER 3

Designing and Analyzing Linear Systems

CONTENTS

3.1 Solution of Square Systems 61
3.1.1 Regression 62
3.1.2 Least Squares 64
3.1.3 Tikhonov Regularization 65
3.1.4 Image Alignment 66
3.1.5 Deconvolution 68
3.1.6 Harmonic Parameterization 69
3.2 Special Properties of Linear Systems 70
3.2.1 Positive Definite Matrices and the Cholesky Factorization 70
3.2.2 Sparsity 74
3.2.3 Additional Special Structures 75
3.3 Sensitivity Analysis 76
3.3.1 Matrix and Vector Norms 76
3.3.2 Condition Numbers 79

NOW that we can solve linear systems of equations, we will show how to apply this generic machinery to solving several practical problems. LU factorization and Gaussian elimination can be applied directly to produce the desired output in each case.

This approach is guaranteed to work, but one natural question is whether there are more efficient or stable algorithms if we know more about the structure of a particular linear problem. Thus, we will examine the matrices A arising in these problems to reveal special properties common to many problems. Writing linear solvers specifically for these classes of matrices will provide speed and numerical advantages at the cost of generality. We also will return to concepts from Chapter 1 to design heuristics evaluating how much we can trust the solution $\vec{x}$ to a linear system $A\vec{x} = \vec{b}$ in the presence of rounding and other sources of error.

3.1 SOLUTION OF SQUARE SYSTEMS

At the beginning of the previous chapter we assumed that we would only consider square, invertible matrices A when solving $A\vec{x} = \vec{b}$. While this restriction was a nontrivial one, many if not most applications of linear solvers can be posed in terms of square, invertible linear systems. We explore a few applications below.

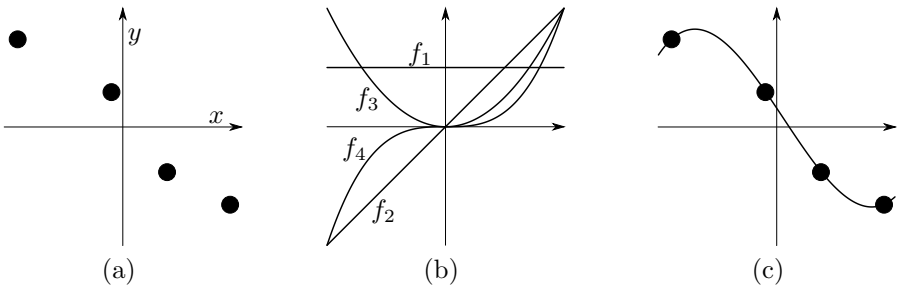


FIGURE 3.1 (a) The input for regression, a set of $(x^{(k)}, y^{(k)})$ pairs; (b) a set of basis functions $\{f_1, f_2, f_3, f_4\}$; (c) the output of regression, a set of coefficients $c_1, \ldots, c_4$ such that the linear combination $\sum_{k=1}^{4} c_k f_k(x)$ goes through the data points.

3.1.1 Regression

We will start with an application from data analysis known as *regression*. Suppose we carry out a scientific experiment and wish to understand the structure of the experimental results. One way to model such a relationship is to write the *independent variables* of a given trial in a vector $\vec{x} \in \mathbb{R}^n$ and to think of the *dependent variable* as a function $f(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$. Given a few $(\vec{x}, f(\vec{x}))$ pairs, our goal is to predict the output of $f(\vec{x})$ for a new $\vec{x}$ without carrying out the full experiment.

Example 3.1 (Biological experiment)**.** Suppose we wish to measure the effects of fertilizer, sunlight, and water on plant growth. We could do a number of experiments applying different amounts of fertilizer (in cm^3), sunlight (in watts), and water (in ml) to a plant and measuring the height of the plant after a few days. Assuming plant height is a direct function of these variables, we can model our observations as samples from a function $f : \mathbb{R}^3 \to \mathbb{R}$ that takes the three parameters we wish to test and outputs the height of the plant at the end of the experimental trial.

In *parametric* regression, we additionally assume that we know the structure of f ahead of time. For example, suppose we assume that f is linear:

$$f(\vec{x}) = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n.$$

Then, our goal becomes more concrete: to estimate the coefficients $a_1, \ldots, a_n$.

Suppose we do a series of experiments that show $\vec{x}^{(k)} \mapsto y^{(k)} \equiv f(\vec{x}^{(k)})$. Continuing the linear example, plugging into the linear form for f shows a set of statements:

$$\begin{aligned} y^{(1)} &= f(\vec{x}^{(1)}) = a_1 x_1^{(1)} + a_2 x_2^{(1)} + \cdots + a_n x_n^{(1)} \\ y^{(2)} &= f(\vec{x}^{(2)}) = a_1 x_1^{(2)} + a_2 x_2^{(2)} + \cdots + a_n x_n^{(2)} \\ &\vdots \end{aligned}$$

Contrary to our earlier notation $A\vec{x} = \vec{b}$, the unknowns here are the a_k's, *not* the $\vec{x}^{(k)}$'s. With this notational difference in mind, if we make exactly n observations we can write:

$$\begin{pmatrix} - & \vec{x}^{(1)\top} & - \\ - & \vec{x}^{(2)\top} & - \\ & \vdots & \\ - & \vec{x}^{(n)\top} & - \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

In other words, if we carry out n trials of our experiment and write the independent variables in the columns of a matrix $X \in \mathbb{R}^{n \times n}$ and the dependent variables in a vector $\vec{y} \in \mathbb{R}^n$, the coefficients $\vec{a}$ can be recovered by solving the linear system $X^\top \vec{a} = \vec{y}$.

We can generalize this method to certain nonlinear forms for the function f using an approach illustrated in Figure 3.1. The key is to write f as a linear combination of *basis functions*. Suppose we assume that $f(\vec{x})$ takes the following form:

$$f(\vec{x}) = a_1 f_1(\vec{x}) + a_2 f_2(\vec{x}) + \cdots + a_m f_m(\vec{x}),$$

where $f_k : \mathbb{R}^n \to \mathbb{R}$ and we wish to estimate the parameters a_k. Then, by a parallel derivation given m observations of the form $\vec{x}^{(k)} \mapsto y^{(k)}$ we can find the parameters by solving:

$$\begin{pmatrix} f_1(\vec{x}^{(1)}) & f_2(\vec{x}^{(1)}) & \cdots & f_m(\vec{x}^{(1)}) \\ f_1(\vec{x}^{(2)}) & f_2(\vec{x}^{(2)}) & \cdots & f_m(\vec{x}^{(2)}) \\ \vdots & \vdots & \cdots & \vdots \\ f_1(\vec{x}^{(m)}) & f_2(\vec{x}^{(m)}) & \cdots & f_m(\vec{x}^{(m)}) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{pmatrix}$$

That is, even if the f's are nonlinear, we can learn weights a_k using purely linear techniques.

Example 3.2 (Linear regression)**.** The system $X^\top \vec{a} = \vec{y}$ from our initial example can be recovered from the general formulation by taking $f_k(\vec{x}) \equiv x_k$.

Example 3.3 (Polynomial regression)**.** As illustrated in Figure 3.1, suppose that we observe a function of a single variable $f(x)$ and wish to write it as an $(n-1)$-st degree polynomial

$$f(x) \equiv a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1} x^{n-1}.$$

Given n pairs $x^{(k)} \mapsto y^{(k)}$, we can solve for the parameters $\vec{a}$ via the system

$$\begin{pmatrix} 1 & x^{(1)} & (x^{(1)})^2 & \cdots & (x^{(1)})^{n-1} \\ 1 & x^{(2)} & (x^{(2)})^2 & \cdots & (x^{(2)})^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x^{(n)} & (x^{(n)})^2 & \cdots & (x^{(n)})^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_2 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}.$$

In other words, we take $f_k(x) = x^{k-1}$ in the general form above. Incidentally, the matrix on the left hand side of this relationship is known as a Vandermonde matrix.

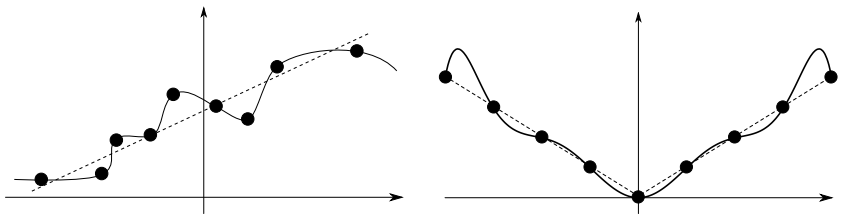
Example 3.4 (Oscillation)**.** A foundational notion from signal processing for audio and images is the decomposition of a function into a linear combination of cosine or sine waves at different frequencies. This decomposition of a function defines its *Fourier transform*.

As the simplest possible case, we can try to recover the parameters of a single-frequency wave. Suppose we wish to find parameters a and ϕ of a function $f(x) = a\cos(x + \phi)$ given two (x, y) samples satisfying $y^{(1)} = f(x^{(1)})$ and $y^{(2)} = f(x^{(2)})$. Although this setup as we have written it is nonlinear, we can recover a and ϕ using the linear method we described above after some mathematical transformations.

From trigonometry, any function of the form $g(x) = a_1 \cos x + a_2 \sin x$ can be written $g(x) = a\cos(x + \phi)$ after applying the formulae

$$a = \sqrt{a_1^2 + a_2^2} \qquad\qquad \phi = -\arctan \frac{a_2}{a_1}.$$

With this change of variables, we can find $f(x)$ by applying the linear method above to



(a) Overfitting (b) Wrong basis

FIGURE 3.2 Drawbacks of fitting function values exactly: (a) noisy data might be better represented by a simple function rather than a complex curve that touches every data point and (b) the basis functions might not be tuned to the function being sampled. In (b), we fit a polynomial of degree eight to nine samples from $f(x) = |x|$ but would have been more successful using a basis of linear segments.

computing the coefficients a_1 and a_2 and then using the formulas above to find a and ϕ. This construction can be extended to fitting functions of the form $f(x) = \sum_k a_k \cos(x + \phi_k)$, giving one way to motivate the discrete Fourier transform of f.

3.1.2 Least Squares

The techniques in §3.1.1 provide valuable methods for finding a continuous f matching a set of data pairs $\vec{x}_k \mapsto y_k$ *exactly*; for this reason they are called *interpolatory* techniques, which we will explore in detail in Chapter 12. They have two related drawbacks, illustrated in Figure 3.2:

- There might be some error in measuring the values $\vec{x}_k$ and y_k. In this case, a simpler $f(\vec{x})$ satisfying the approximate relationship $f(\vec{x}_k) \approx y_k$ may be acceptable or even preferable to an exact $f(\vec{x}_k) = y_k$ that goes through each data point.
- If there are m functions $f_1, \ldots, f_m$, then we need exactly m observations $\vec{x}_k \mapsto y_k$. Additional observations would have to be thrown out, or we would have to introduce more f_k's, which can become increasingly complicated.

Both of these issues are related to the larger problem of *over-fitting*: Fitting a function with n degrees of freedom to n data points leaves no room for measurement error.

More generally, suppose we wish to solve the linear system $A\vec{x} = \vec{b}$ for $\vec{x}$. If we denote row k of A as $\vec{r}_k^\top$, then our system looks like

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} - & \vec{r}_1^\top & - \\ - & \vec{r}_2^\top & - \\ \vdots & \vdots & \vdots \\ - & \vec{r}_n^\top & - \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ by expanding } A\vec{x}$$

$$= \begin{pmatrix} \vec{r}_1 \cdot \vec{x} \\ \vec{r}_2 \cdot \vec{x} \\ \vdots \\ \vec{r}_n \cdot \vec{x} \end{pmatrix} \text{ by definition of matrix multiplication.}$$

From this perspective, each row of the system corresponds to a separate observation of the form $\vec{r}_k \cdot \vec{x} = b_k$. That is, an alternative way to interpret the linear system $A\vec{x} = \vec{b}$ is that it encodes n statements of the form, "The dot product of $\vec{x}$ with $\vec{r}_k$ is b_k."

From this viewpoint, a tall system $A\vec{x} = \vec{b}$ where $A \in \mathbb{R}^{m\times n}$ and $m > n$ encodes more than n of these dot product observations. When we make more than n observations, however, they may be *incompatible*; as explained §2.1, tall systems do not have to admit a solution.

When we cannot solve $A\vec{x} = \vec{b}$ exactly, we can relax the problem and try to find an approximate solution $\vec{x}$ satisfying $A\vec{x} \approx \vec{b}$. One of the most common ways to solve this problem, known as *least-squares*, is to ask that the residual $\vec{b} - A\vec{x}$ be as small as possible by minimizing the norm $\|\vec{b} - A\vec{x}\|_2$. If there is an exact solution $\vec{x}$ satisfying the tall system $A\vec{x} = \vec{b}$, then the minimum of this energy is zero, since norms are nonnegative and in this case $\|\vec{b} - A\vec{x}\|_2 = \|\vec{b} - \vec{b}\|_2 = 0$.

Minimizing $\|\vec{b} - A\vec{x}\|_2$ is the same as minimizing $\|\vec{b} - A\vec{x}\|_2^2$, which we expanded in Example 0.16 to:

$$\|\vec{b} - A\vec{x}\|_2^2 = \vec{x}^\top A^\top A\vec{x} - 2\vec{b}^\top A\vec{x} + \|\vec{b}\|_2^2.^*$$

The gradient of this expression with respect to $\vec{x}$ must be zero at its minimum, yielding the following system:

$$\vec{0} = 2A^\top A\vec{x} - 2A^\top\vec{b}$$

$$\text{Or equivalently:} \qquad A^\top A\vec{x} = A^\top\vec{b}.$$

This famous relationship is worthy of a theorem:

Theorem 3.1 (Normal equations)**.** Minima of the residual $\|\vec{b} - A\vec{x}\|_2$ for $A \in \mathbb{R}^{m\times n}$ (with no restriction on m or n) satisfy $A^\top A\vec{x} = A^\top\vec{b}$.

If at least n rows of A are linearly independent, then the matrix $A^\top A \in \mathbb{R}^{n\times n}$ is invertible. In this case, the minimum residual occurs uniquely at $(A^\top A)^{-1}A^\top\vec{b}$. Put another way:

In the overdetermined case, solving the least-squares problem $A\vec{x} \approx \vec{b}$ is equivalent to solving the *square* linear system $A^\top A\vec{x} = A^\top\vec{b}$.

With this construction, we have expanded our set of solution strategies to tall $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ by applying only techniques for square matrices.

3.1.3 Tikhonov Regularization

When solving linear systems, the underdetermined case $m < n$ is considerably more difficult to handle due to increased ambiguity. As discussed in §2.1, in this case we lose the possibility of a *unique* solution to $A\vec{x} = \vec{b}$. To choose between the possible solutions, we must make an additional assumption on $\vec{x}$ to obtain a unique solution, e.g. that it has a small norm or that it contains many zeros. Each such *regularizing* assumption leads to a different solution strategy. The particular choice of a regularizer may be application-dependent, but here we outline one general approach commonly applied in statistics and machine learning; we will introduce an alternative in §6.2.1 after introducing the SVD of a matrix.

*If this result is not intuitive, it may be valuable to return to the material in §0.4 at this point for review.

When there are multiple vectors $\vec{x}$ that all minimize $\|A\vec{x}-\vec{b}\|_2^2$, as in the underdetermined case discussed above, we might say that the least-square energy function is *insufficient* in the sense that it does not admit a unique minimizer. For this reason, for fixed $\alpha > 0$, we might introduce an additional term to our minimization problem:

$$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2 + \alpha\|\vec{x}\|_2^2$$

This second term is known as a *Tikhonov regularizer*. When $0 < \alpha \ll 1$, this optimization effectively asks that among the minimizers of $\|A\vec{x} - \vec{b}\|_2$ we would prefer those with small norm $\|\vec{x}\|_2$; as α increases, we prioritize the norm of $\vec{x}$ more. This energy is the product of an "Occam's razor" philosophy: In the absence of more information about $\vec{x}$ we might as well choose an $\vec{x}$ with small entries.

To minimize this new objective, we take the derivative with respect to $\vec{x}$ and set it equal to zero:

$$\vec{0} = 2A^\top A\vec{x} - 2A^\top\vec{b} + 2\alpha\vec{x}$$

Or, equivalently:

$$(A^\top A + \alpha I_{n\times n})\vec{x} = A^\top\vec{b}$$

So, if we wish to introduce Tikhonov regularization to a linear problem, all we have to do is add α down the diagonal of the least-squares matrix $A^\top A$.

When $A\vec{x} = \vec{b}$ is underdetermined, the matrix $A^\top A$ is not invertible. The new Tikhonov term, however, solves this issue, since for $\vec{x} \neq \vec{0}$:

$$\vec{x}^\top(A^\top A + \alpha I_{n\times n})\vec{x} = \|A\vec{x}\|_2^2 + \alpha\|\vec{x}\|_2^2 > 0.$$

The strict $>$ holds because $\vec{x} \neq \vec{0}$ and implies that $A^\top A + \alpha I_{n\times n}$ cannot have a null space vector $\vec{x}$. Hence, regardless of A the Tikhonov-regularized system of equations is invertible; in the language we will introduce in §3.2.1, it is *positive definite*.

Tikhonov regularization is effective for dealing with null spaces and numerical issues. When A is poorly-conditioned, adding this type of regularization can improve conditioning even when the original system was solvable. We acknowledge two drawbacks, however, that can require more advanced methods when they appear:

- The solution $\vec{x}$ of the Tikhonov-regularized system no longer satisfies $A\vec{x} = \vec{b}$ exactly.
- When α is small, the matrix $A^\top A+\alpha I_{n\times n}$ is invertible but may be poorly conditioned. Increasing α solves this problem at the cost of less accurate solutions to $A\vec{x} = \vec{b}$.

When the columns of A span $\mathbb{R}^m$, an alternative to Tikhonov regularization is to minimize $\|\vec{x}\|_2$ with the "hard" constraint $A\vec{x} = \vec{b}$. Problem 3.5 shows that this least-norm solution is given by $\vec{x} = A^\top(AA^\top)^{-1}\vec{b}$, a similar formula to the normal equations for least-squares.

3.1.4 Image Alignment

Suppose we take two photographs of the same scene from different positions. One common task in computer vision and graphics is to stitch them together to make a single larger image. To do so, the user (or an automatic system) marks p pairs of points $\vec{x}_k, \vec{y}_k \in \mathbb{R}^2$ such that for each k the location $\vec{x}_k$ in image one corresponds to the location $\vec{y}_k$ in image two. Then, the software automatically warps the second image onto the first or vice versa such that the pairs of points are aligned.



(a) (b) Input images with keypoints (c) Aligned images

FIGURE 3.3 (a) The image alignment problem attempts to find the parameters A and $\vec{b}$ of a transformation from one image of a scene to another using labeled keypoints $\vec{x}$ on the first image paired with points $\vec{y}$ on the second. As an example, keypoints marked in white on the two images in (b) are used to create the aligned image (c).

When the camera makes a small motion, a reasonable assumption is that there exists some transformation matrix $A \in \mathbb{R}^{2\times 2}$ and a translation vector $\vec{b} \in \mathbb{R}^2$ such that for all k,

$$\vec{y}_k \approx A\vec{x}_k + \vec{b}.$$

That is, position $\vec{x}$ on image one should correspond to position $A\vec{x} + \vec{b}$ on image two. Figure 3.3(a) illustrates this notation.

With this assumption, given a set of pairs of corresponding points $(\vec{x}_1, \vec{y}_1), \ldots, (\vec{x}_p, \vec{y}_p)$, our goal is to compute A and $\vec{b}$ matching these points as closely as possible. Beyond numerical issues, mistakes may have been made while locating the corresponding points, however, and we must account for approximation error due to the slightly-nonlinear camera projection of real-world lenses. Thus, rather than requiring that the marked points match exactly we can ask that they are matched in a least-squares sense. In other words, we solve the following minimization problem:

$$\min_{A,\vec{b}} \sum_{k=1}^{p} \|(A\vec{x}_k + \vec{b}) - \vec{y}_k\|_2^2.$$

This problem has six unknowns total, the four elements of A and the two elements of $\vec{b}$. Figure 3.3(b,c) shows typical output for this method; five keypoints rather than the required three are used to stabilize the output transformation using least-squares.

This expression is a sum of squared linear expressions in our unknowns A and $\vec{b}$, and as we will see it can be solved linearly. Define

$$f(A, \vec{b}) = \sum_k \|(A\vec{x}_k + \vec{b}) - \vec{y}_k\|_2^2.$$

Then, we can simplify f as follows:

$$\begin{aligned} f(A, \vec{b}) &= \sum_k (A\vec{x}_k + \vec{b} - \vec{y}_k)^\top (A\vec{x}_k + \vec{b} - \vec{y}_k) \text{ since } \|\vec{v}\|_2^2 = \vec{v}^\top \vec{v} \\ &= \sum_k \left[\vec{x}_k^\top A^\top A\vec{x}_k + 2\vec{x}_k^\top A^\top \vec{b} - 2\vec{x}_k^\top A^\top \vec{y}_k + \vec{b}^\top \vec{b} - 2\vec{b}^\top \vec{y}_k + \vec{y}_k^\top \vec{y}_k\right] \\ &\qquad \text{where terms with leading 2 apply the fact } \vec{a}^\top \vec{b} = \vec{b}^\top \vec{a} \end{aligned}$$



(a) Sharp (b) Blurry (c) Deconvolved (d) Difference

FIGURE 3.4 Suppose rather than taking the sharp image (a) we accidentally take a blurry photo (b); then, deconvolution can be used to recover a sharp approximation of the original image (c). The difference between (a) and (c) is shown in (d); only high-frequency detail is different between the two images.

To find where f is minimized, we differentiate with respect to both $\vec{b}$ and with respect to the elements of A, and set both of the resulting expressions equal to zero. This method leads to the following system:

$$0 = \nabla_{\vec{b}} f(A, \vec{b}) = \sum_k \left[2A\vec{x}_k + 2\vec{b} - 2\vec{y}_k \right]$$

$$0 = \nabla_A f(A, \vec{b}) = \sum_k \left[2A\vec{x}_k\vec{x}_k^\top + 2\vec{b}\vec{x}_k^\top - 2\vec{y}_k\vec{x}_k^\top \right] \text{ by the identities in Problem 3.1}$$

In the second equation, we denote the gradient $\nabla_A f$ as the *matrix* whose entries are $\partial f / \partial A_{ij}$. Simplifying somewhat, if we define $X \equiv \sum_k \vec{x}_k\vec{x}_k^\top$, $\vec{x}_{\text{sum}} \equiv \sum_k \vec{x}_k$, $\vec{y}_{\text{sum}} \equiv \sum_k \vec{y}_k$, and $C \equiv \sum_k \vec{y}_k\vec{x}_k^\top$, then the optimal A and $\vec{b}$ satisfy the linear system:

$$A\vec{x}_{\text{sum}} + p\vec{b} = \vec{y}_{\text{sum}}$$
$$AX + \vec{b}\vec{x}_{\text{sum}}^\top = C$$

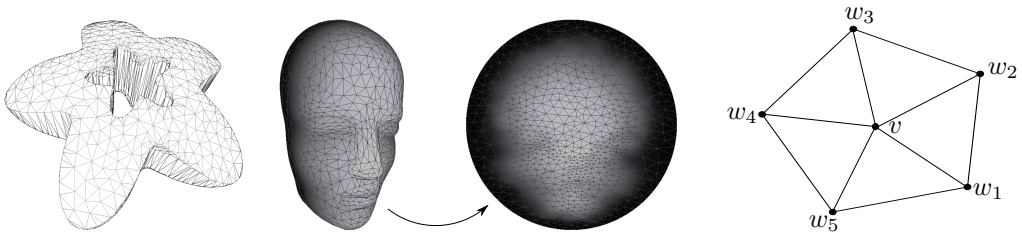
This system is linear in the unknowns A and $\vec{b}$; in Problem 3.2 you will write the 6×6 system explicitly.

This example illustrates a larger pattern in modeling least-squares problems. We started by defining a simple relationship between the unknowns that we wanted to be true, namely $(A\vec{x} + \vec{b}) - \vec{y} \approx \vec{0}$. Given a number of data points $(\vec{x}_k, \vec{y}_k)$, we designed an energy measuring the quality of a choice of A and $\vec{b}$ by summing up the squared norms of terms we wished to be zero: $\sum_k \|(A\vec{x}_k + \vec{b}) - \vec{y}_k\|_2^2$. Differentiating this sum with respect to the unknowns gave a *linear* system of equations to solve for the best possible choice. This pattern is a common source of optimization problems that can be solved linearly.

3.1.5 Deconvolution

An artist hastily taking pictures of a scene may accidentally take photographs that are slightly out of focus. While a photo that is completely blurred may be a lost cause, if there is only localized or small-scale blurring, we may be able to recover a sharper image using computational techniques. One strategy is *deconvolution*, explained below; an example test case of the method outlined below is shown in Figure 3.4.

We can think of a grayscale photograph as a point in $\mathbb{R}^p$, where p is the number of pixels



(a) Triangle mesh (b) Parameterization (c) Harmonic condition

FIGURE 3.5 (a) An example of a triangle mesh, the typical structure used to represent three-dimensional shapes in computer graphics. (b) In mesh parameterization, we seek a map from a three-dimensional mesh (left) to the two-dimensional image plane (right); the right-hand side shown here was computed using the method suggested in §3.1.6. (c) The harmonic condition is that the position of vertex v is the average of the positions of its neighbors $w_1, \ldots, w_5$.

it contains; each pixel's intensity is stored in a different dimension. If the photo is in color we may need red, green, and blue intensities per pixel, yielding a similar representation in $\mathbb{R}^{3p}$. Regardless, most image blurs are *linear*, e.g. Gaussian convolution or operations averaging a pixel's intensity with those of its neighbors. In image processing, these linear operations can be encoded using a matrix G taking a sharp image $\vec{x}$ to its blurred counterpart $G\vec{x}$.

Suppose we take a blurry photo $\vec{x}_0 \in \mathbb{R}^p$. Then, we could try to recover the underlying sharp image $\vec{x} \in \mathbb{R}^p$ by solving the least-squares problem

$$\min_{\vec{x}\in\mathbb{R}^p} \|\vec{x}_0 - G\vec{x}\|_2^2.$$

This minimization assumes that when you blur $\vec{x}$ with G, you get the observed photo $\vec{x}_0$. By the same construction as previous chapters, if we know G then this problem can be solved using linear methods.

In practice, this optimization might be unstable since it is solving a difficult inverse problem. In particular, many pairs of distinct images look very similar after they are blurred, making the reverse operation more difficult. One resolution is to use Tikhonov regularization:

$$\min_{\vec{x}\in\mathbb{R}^p} \|\vec{x}_0 - G\vec{x}\|_2^2 + \alpha\|\vec{x}\|_2^2$$

More complex models may constrain $\vec{x} \geq 0$, since negative intensities are not reasonable, but adding such a constraint makes the optimization nonlinear and better solved by the methods we will introduce starting in Chapter 9.

3.1.6 Harmonic Parameterization

Systems for animation often represent geometric objects in a scene using *triangle meshes*, sets of points linked together into triangles as in Figure 3.5(a). To give these meshes fine textures and visual detail, a common practice is to store a detailed color texture as an image or photograph, and to map this texture onto the geometry. Each vertex of the mesh then carries not only its geometric location in space but also *texture coordinates* representing its position on the texture plane.

Mathematically, a mesh can be represented as a collection of n vertices $V \equiv \{v_1, \ldots, v_n\}$

linked in pairs by edges $E \subseteq V \times V$. Geometrically, each vertex $v \in V$ is associated with a location $\vec{x}(v)$ in three-dimensional space $\mathbb{R}^3$. Additionally, we will decorate each vertex with a texture coordinate $\vec{t}(v) \in \mathbb{R}^2$ describing its location in the image plane; it is desirable for these positions to be laid out smoothly to avoid squeezing or stretching the texture relative to the geometry of the surface. With this criterion in mind, the problem of *parameterization* is to fill in the positions $\vec{t}(v)$ for all the vertices $v \in V$ given a few positions laid out manually; desirable mesh parameterizations minimize the geometric distortion of the mesh from its configuration in three-dimensional space to the plane. Surprisingly, many state-of-the-art parameterization algorithms involve little more than a linear solve; we will outline one method originally proposed in [117].

For simplicity, suppose that the mesh has disk topology, meaning that it can be mapped to the interior of a circle in the plane, and that we have fixed the location of each vertex on its boundary $B \subseteq V$. The job of the parameterization algorithm then is to fill in positions for the interior vertices of the mesh. This setup and the output of the algorithm outlined below are shown in Figure 3.5(b).

For a vertex $v \in V$, take $N(v)$ to be the set of neighbors of v on the mesh, given by

$$N(v) \equiv \{w \in V : (v, w) \in E\}.$$

Then, for each vertex $v \in V \backslash B$, a reasonable criterion for parameterization quality is that v should be located at the center of its neighbors, illustrated in Figure 3.5(c). Mathematically, this condition can be expressed as:

$$\vec{t}(v) = \frac{1}{|N(v)|} \sum_{w \in N(v)} \vec{t}(w)$$

Then, we can associate each $v \in V$ with a linear condition either fixing its position on the boundary or asking that its position equal the average of its neighboring positions. This $|V| \times |V|$ system of equations generates a *harmonic* parameterization.

The final output in Figure 3.5(b) is laid out in an elastic fashion that evenly distributes its vertices on the image plane. Harmonic parameterization has been extended in numerous ways, e.g. by accounting for the lengths of the edges in E as they are realized in three-dimensional space.

3.2 SPECIAL PROPERTIES OF LINEAR SYSTEMS

The examples above have provided several contexts in which linear systems of equations can be used to solve practical computing problems. As derived in the previous chapter, Gaussian elimination can be used to solve all of these problems in polynomial time, but it remains to be seen whether they can be solved using faster or stabler techniques. With this question in mind, here we look more closely at the matrices from §3.1 to reveal that they have many properties in common; then, by deriving solution techniques for these special classes of matrices we will achieve better speed and numerical quality on these particular problems.

3.2.1 Positive Definite Matrices and the Cholesky Factorization

As shown in Theorem 3.1, solving a least-squares problem $A\vec{x} \approx \vec{b}$ yields a solution $\vec{x}$ satisfying the square linear system $(A^\top A)\vec{x} = A^\top \vec{b}$. Regardless of A, the matrix $A^\top A$ has a few special properties that make this system special.

First, $A^\top A$ is symmetric, as can be shown by applying the identities $(AB)^\top = B^\top A^\top$ and $(A^\top)^\top = A$:

$$(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A.$$

We can express this symmetry index-wise by writing $(A^\top A)_{ij} = (A^\top A)_{ji}$ for all indices i, j. This property implies that it is sufficient to store only the values of $A^\top A$ on or above the diagonal, since the rest of the elements can be obtained by symmetry.

Furthermore, $A^\top A$ is a *positive semidefinite* matrix, as defined below:

Definition 3.1 (Positive (Semi-)Definite)**.** A matrix $B \in \mathbb{R}^{n \times n}$ is positive semidefinite if for all $\vec{x} \in \mathbb{R}^n$, $\vec{x}^\top B \vec{x} \geq 0$. B is positive definite if $\vec{x}^\top B \vec{x} > 0$ whenever $\vec{x} \neq \vec{0}$.

We can characterize the definiteness of $A^\top A$ using an observation:

Proposition 3.1. For any $A \in \mathbb{R}^{m \times n}$, the matrix $A^\top A$ is positive semidefinite. Furthermore, $A^\top A$ is positive definite exactly when the columns of A are linearly independent.

Proof. We first check that $A^\top A$ is *always* positive semidefinite. To do so, take any $\vec{x} \in \mathbb{R}^n$. Then,

$$\vec{x}^\top (A^\top A)\vec{x} = (A\vec{x})^\top (A\vec{x}) = (A\vec{x}) \cdot (A\vec{x}) = \|A\vec{x}\|_2^2 \geq 0.$$

To prove the second statement, first suppose the columns of A are linearly independent. If A were only semidefinite then there would be an $\vec{x} \neq \vec{0}$ with $\vec{x}^\top A^\top A \vec{x} = 0$, but as shown above this would imply $\|A\vec{x}\|_2 = 0$, or equivalently $A\vec{x} = \vec{0}$, contradicting the independence of the columns of A. Conversely, if A has linearly *dependent* columns, then there exists a $\vec{y} \neq \vec{0}$ with $A\vec{y} = \vec{0}$, so then $\vec{y}^\top A^\top A \vec{y} = \vec{0}^\top \vec{0} = 0$, and hence A is not positive definite. □

As a corollary, $A^\top A$ is invertible exactly when A has linearly independent columns, providing a condition to check to verify whether a least-squares problem admits a unique solution.

Given the prevalence of the least-squares system $A^\top A\vec{x} = A^\top \vec{b}$, it is worth considering the possibility of writing faster linear solvers specially-designed for this case. In particular, suppose we wish to solve a *symmetric positive definite* system $C\vec{x} = \vec{d}$; from our discussion we could take $C = A^\top A$ and $\vec{d} = A^\top \vec{b}$, although there exist many systems that naturally are symmetric and positive definite without explicitly coming from a least-squares model. As we have already explored, we could LU-factorize the matrix C to solve the system using Gaussian elimination, but we can do somewhat better.

Aside 3.1 (Block matrix notation)**.** Our construction in this section will rely on *block matrix* notation. This notation builds larger matrices out of smaller ones. For example, suppose $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{m \times k}$, $C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{p \times k}$. Then, we could construct a larger matrix by writing:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathbb{R}^{(m+p) \times (n+k)}$$

This "block matrix" is constructed by concatenation. Block matrix notation is convenient, but we must be careful to concatenate matrices with dimensions that match. Also, the mechanisms of matrix algebra extend to this case, e.g.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} AE + BG & AF + BH \\ CE + DG & CF + DH \end{pmatrix}$$

We will proceed without checking these identities explicitly, but as an exercise it is worth double-checking that they are true.

We can deconstruct our symmetric positive-definite $C \in \mathbb{R}^{n \times n}$ as a block matrix:

$$C = \begin{pmatrix} c_{11} & \vec{v}^\top \\ \vec{v} & \tilde{C} \end{pmatrix}$$

where $\vec{v} \in \mathbb{R}^{n-1}$ and $\tilde{C} \in \mathbb{R}^{(n-1)\times(n-1)}$. Thanks to the special structure of C, we can make the following observation:

$$\begin{aligned} \vec{e}_1^\top C \vec{e}_1 &= \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} c_{11} & \vec{v}^\top \\ \vec{v} & \tilde{C} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} c_{11} \\ \vec{v} \end{pmatrix} \\ &= c_{11} \\ &> 0 \text{ since } C \text{ is positive definite and } \vec{e}_1 \neq \vec{0}. \end{aligned}$$

This argument shows that we do not have to use pivoting to guarantee that $c_{11} \neq 0$ in the first step of Gaussian elimination.

Continuing with Gaussian elimination, we can apply a forward substitution matrix E of the form

$$E = \begin{pmatrix} 1/\sqrt{c_{11}} & \vec{0}^\top \\ \vec{r} & I_{(n-1)\times(n-1)} \end{pmatrix}.$$

Here, the vector $\vec{r} \in \mathbb{R}^{n-1}$ contains forward substitution scaling factors such that $r_{i-1}c_{11} = -c_{i1}$. Unlike our original construction of Gaussian elimination, we scale row 1 by $1/\sqrt{c_{11}}$ for reasons that will become apparent shortly.

By design, after forward substitution we know the form of the product EC to be:

$$EC = \begin{pmatrix} \sqrt{c_{11}} & \vec{v}^\top/\sqrt{c_{11}} \\ \vec{0} & D \end{pmatrix}$$

for some $D \in \mathbb{R}^{(n-1)\times(n-1)}$.

Now we diverge from the derivation of Gaussian elimination: Rather than moving on to the second row, to maintain symmetry we can post-multiply by $E^\top$ to obtain a product $ECE^\top$:

$$\begin{aligned} ECE^\top &= (EC)E^\top \\ &= \begin{pmatrix} \sqrt{c_{11}} & \vec{v}^\top/\sqrt{c_{11}} \\ \vec{0} & D \end{pmatrix} \begin{pmatrix} 1/\sqrt{c_{11}} & \vec{r}^\top \\ \vec{0} & I_{(n-1)\times(n-1)} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \vec{0}^\top \\ \vec{0} & D \end{pmatrix} \end{aligned}$$

The $\vec{0}^\top$ in the upper right follows from construction of E as an elimination matrix. Alternatively, an easier if less direct argument is that $ECE^\top$ is symmetric, and the lower-left element of the block form for $ECE^\top$ is $\vec{0}$ by block matrix multiplication. Regardless, we have eliminated the first row *and* the first column of C! Furthermore, the remaining submatrix D is also positive definite.

We can repeat this process to eliminate all the rows and columns of C symmetrically. This solution is *specific* to symmetric positive-definite matrices, since

- symmetry allowed us to apply the same E to both sides, and
- positive definiteness guaranteed that $c_{11} > 0$, thus implying that $1/\sqrt{c_{11}}$ exists.

Similar to LU factorization we now obtain a factorization $C = LL^\top$ for a lower triangular matrix L. This factorization is constructed by applying elimination matrices symmetrically using the process above, until we reach

$$E_k \cdots E_2 E_1 C E_1^\top E_2^\top \cdots E_k^\top = I_{n\times n}.$$

Then, like our construction in §2.5.1, we define L as a product of lower triangular matrices:

$$L \equiv E_1^{-1} E_2^{-1} \cdots E_k^{-1}.$$

The product $C = LL^\top$ is known as the *Cholesky factorization* of C. If taking the square roots along the diagonal causes numerical issues, a related $LDL^\top$ factorization, where D is a diagonal matrix, avoids this issue and can be derived from the discussion above.

The Cholesky factorization has many practical properties. It takes half the memory to store L from the Cholesky factorization rather than the LU factorization of C. Specifically, L has $n(n+1)/2$ nonzero elements, while the compressed storage of LU factorizations explained in §2.5.3 requires n^2 nonzeros. Furthermore, as with the LU decomposition, solving $C\vec{x} = \vec{d}$ can be accomplished using fast forward and back substitution. Finally, the product $LL^\top$ is symmetric and positive semidefinite regardless of L; if we factored $C = LU$ but made rounding and other mistakes, in degenerate cases the product $C \approx LU$ may no longer satisfy these criteria exactly.

Code for Cholesky factorization can be very succinct. To derive a particularly compact form, we can work backward from the factorization $C = LL^\top$ now that such an object exists. Suppose we choose an arbitrary $k \in \{1, \ldots, n\}$ and write L in block form isolating the k-th row and column:

$$L = \begin{pmatrix} L_{11} & \vec{0} & 0 \\ \vec{\ell}_k^\top & \ell_{kk} & \vec{0}^\top \\ L_{31} & \vec{\ell}_k' & L_{33} \end{pmatrix}$$

Here, since L is lower-triangular, L_{11} and L_{33} are both lower triangular square matrices. Then, applying block matrix algebra to the product $C = LL^\top$ shows:

$$\begin{aligned} C = LL^\top &= \begin{pmatrix} L_{11} & \vec{0} & 0 \\ \vec{\ell}_k^\top & \ell_{kk} & \vec{0}^\top \\ L_{31} & \vec{\ell}_k' & L_{33} \end{pmatrix} \begin{pmatrix} L_{11}^\top & \vec{\ell}_k & L_{31}^\top \\ \vec{0}^\top & \ell_{kk} & (\vec{\ell}_k')^\top \\ 0 & \vec{0} & L_{33}^\top \end{pmatrix} \\ &= \begin{pmatrix} \times & \times & \times \\ \vec{\ell}_k^\top L_{11}^\top & \vec{\ell}_k^\top \vec{\ell}_k + \ell_{kk}^2 & \times \\ \times & \times & \times \end{pmatrix} \end{aligned}$$

We leave out values of the product that are not necessary for our derivation.

Returning to the relationship $C = LL^\top$, from the product above we now have $c_{kk} = \vec{\ell}_k^\top \vec{\ell}_k + \ell_{kk}^2$, or equivalently:

$$\ell_{kk} = \sqrt{c_{kk} - \|\vec{\ell}_k\|_2^2}.$$

where $\vec{\ell}_k \in \mathbb{R}^{k-1}$ contains the elements of the k-th row of L to the left of the diagonal. We can choose $\ell_{kk} \geq 0$ since scaling columns of L by -1 has no effect on the factorization $C = LL^\top$. Furthermore, the middle left element of the product shows

$$L_{11}\vec{\ell}_k = \vec{c}_k$$

```
function CHOLESKY-FACTORIZATION(C)
  ▷ Factors C = LL^T, assuming C is symmetric and positive definite
  L ← C                                  ▷ This algorithm destructively replaces C with L
  for k ← 1, 2, ..., n
    ▷ Back-substitute to place ℓ_k^⊤ at the beginning of row k
    for i ← 1, ..., k − 1                ▷ Current element i of ℓ_k
      s ← 0
      ▷ Iterate over L_11; j < i, so the iteration maintains L_kj = (ℓ_k)_j.
      for j ← 1, ..., i − 1 : s ← s + L_ij L_kj
      L_ki ← (L_ki − s)/L_ii

    ▷ Apply the formula for ℓ_kk
    v ← 0                                ▷ For computing ||ℓ_k||_2^2
    for j ← 1, ..., i − 1 : v ← v + L_kj^2
    L_kk ← sqrt(L_kk − v)
  return L
```

FIGURE 3.6 Cholesky factorization for writing $C = LL^\top$, where the input C is symmetric and positive-definite and the output L is lower-triangular.

where $\vec{c}_k$ contains the elements of C in the same position as $\vec{\ell}_k$. Since L_{11} is lower triangular, this system can be solved by forward substitution for $\vec{\ell}_k$!

Synthesizing the formulas above reveals an algorithm for computing the Cholesky factorization by iterating $k = 1, 2, \ldots, n$. In each step, L_{11} will already be computed by the time we reach row k, giving a way to find $\vec{\ell}_k$ via substitution; similarly ℓ_{kk} comes from the square root formula. We provide pseudocode in Figure 3.6. As with LU factorization, this algorithm runs in $O(n^3)$ time; more specifically, Cholesky factorization takes approximately $\frac{1}{3}n^3$ operations, half the work needed for LU.

3.2.2 Sparsity

We set out in this section to identify properties of specific linear systems that can make them solvable using more efficient techniques than Gaussian elimination. In addition to positive definiteness, many linear systems of equations naturally enjoy *sparsity* properties, meaning that most of the entries of A in the system $A\vec{x} = \vec{b}$ are exactly zero. Sparsity can reflect particular structure in a given problem, including the following use cases:

- In image processing (e.g. §3.1.5), systems for photo editing model using relationships between the values of pixels and those of their neighbors on the image grid. An image may be a point in $\mathbb{R}^p$ for p pixels, but when solving $A\vec{x} = \vec{b}$ for a new size-p image, $A \in \mathbb{R}^{p\times p}$ may have only $O(p)$ rather than $O(p^2)$ nonzeros since each row only involves a single pixel and its up/down/left/right neighbors.
- In computational geometry (e.g. §3.1.6), shapes are often expressed using collections of triangles linked together into a mesh. Equations for surface smoothing, parameterization, and other tasks link values associated with given vertex with only those at their neighbors in the mesh.
- In machine learning, a *graphical model* uses a graph structure $G \equiv (V, E)$ to express

probability distributions over several variables. Each variable is represented using a node $v \in V$ of the graph, and edge $e \in E$ represents a probabilistic dependence. Linear systems arising in this context often have one row per vertex $v \in V$ with nonzeros only in columns involving v and its neighbors.

If $A \in \mathbb{R}^{n \times n}$ is sparse to the point that it contains $O(n)$ rather than $O(n^2)$ nonzero values, there is no reason to store A with n^2 values. Instead, *sparse matrix* storage techniques only store the $O(n)$ nonzeros in a more reasonable data structure, e.g. a list of row/column/value triplets. The choice of a matrix data structure involves considering the likely operations that will occur on the matrix, possibly including multiplication, iteration over nonzeros, or iterating over individual rows or columns.

Unfortunately, the LU and Cholesky factorization of a sparse A may not result in sparse L and U matrices; this loss of structure severely limits the applicability of using these methods to solve $A\vec{x} = \vec{b}$ when A is large but sparse. Thankfully, there are many *direct sparse solvers* adapting LU to sparse matrices that can produce an LU-like factorization without inducing much *fill*, or additional nonzeros; discussion of these techniques is outside the scope of this text. Alternatively, *iterative* techniques have been used to obtain approximate solutions to linear systems; see Chapter 10 for several examples.

3.2.3 Additional Special Structures

Certain matrices are not only sparse but also *structured*. For instance, a *tridiagonal* system of linear equations has the following pattern of nonzero values:

$$\begin{pmatrix} \times & \times & & & \\ \times & \times & \times & & \\ & \times & \times & \times & \\ & & \times & \times & \times \\ & & & \times & \times \end{pmatrix}$$

In the exercises following this chapter, you will derive a special version of Gaussian elimination for dealing with this this *banded* structure.

In other cases, matrices may not be sparse but might admit a sparse representation. For example, consider the *cyclic* matrix:

$$\begin{pmatrix} a & b & c & d \\ d & a & b & c \\ c & d & a & b \\ b & c & d & a \end{pmatrix}$$

This matrix can be stored using only the values a, b, c, d. Specialized techniques for this and other classes of matrices are well-studied and often more efficient than generic Gaussian elimination.

Broadly speaking, once a problem has been reduced to a linear system $A\vec{x} = \vec{b}$, Gaussian elimination provides only one option for how to find $\vec{x}$. It may be possible to show that the matrix A for the given problem can be solved more easily by identifying common special properties such as positive-definiteness, sparsity, and so on. This additional knowledge about A can uncover higher-quality specialized solution techniques. Interested readers should refer to the discussion in [46] for consideration of numerous cases like the ones above.

3.3 SENSITIVITY ANALYSIS

As we have seen, it is important to examine the matrix of a linear system to find out if it has special properties that can simplify the solution process. Sparsity, positive definiteness, symmetry, and so on all provide clues to the proper algorithm to use for a particular problem. Even if a given solution strategy might work in theory, however, it is equally important to understand how well we can trust the output. For instance, due to rounding and other discrete effects, it might be the case that an implementation of Gaussian elimination for solving $A\vec{x} = \vec{b}$ yields a solution $\vec{x}_0$ such that $0 < \|A\vec{x}_0 - \vec{b}\| \ll 1$; in other words, $\vec{x}_0$ only solves the system approximately.

One general way to understand the likelihood of error is through *sensitivity analysis*. To measure sensitivity, we ask what might happen to $\vec{x}$ if instead of solving $A\vec{x} = \vec{b}$ in reality we solve a *perturbed* system of equations $(A + \delta A)\vec{x} = \vec{b} + \delta\vec{b}$. There are two ways of viewing conclusions made by this type of analysis:

1. We may represent A and $\vec{b}$ inexactly thanks to rounding and other effects. This analysis then shows the best possible accuracy we can expect for $\vec{x}$ given the mistakes made representing the problem.

2. Suppose our solver generates an inexact approximation $\vec{x}_0$ to the solution $\vec{x}$ of $A\vec{x} = \vec{b}$. This vector $\vec{x}_0$ itself is the exact solution of a different system $A\vec{x}_0 = \vec{b}_0$ if we *define* $\vec{b}_0 \equiv A\vec{x}_0$ (be sure you understand why this sentence is not a tautology!). Understanding how changes in $\vec{x}_0$ affect changes in $\vec{b}_0$ show how sensitive the system is to slightly incorrect answers.

Our discussion here is similar to and indeed motivated by our definitions of forward and backward error in §1.2.1.

3.3.1 Matrix and Vector Norms

Before we can discuss the sensitivity of a linear system, we have to be somewhat careful to define what it means for a change $\delta\vec{x}$ to be "small." Generally, we wish to measure the length, or *norm*, of a vector $\vec{x}$. We have already encountered the two-norm of a vector:

$$\|\vec{x}\|_2 \equiv \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

for $\vec{x} \in \mathbb{R}^n$. This norm is popular thanks to its connection to Euclidean geometry, but it is by no means the only norm on $\mathbb{R}^n$. Most generally, we define a *norm* as follows:

Definition 3.2 (Vector norm)**.** A vector norm is a function $\|\cdot\| : \mathbb{R}^n \to [0, \infty)$ satisfying the following conditions:

- $\|\vec{x}\| = 0$ if and only if $\vec{x} = \vec{0}$ ("$\|\cdot\|$ separates points")
- $\|c\vec{x}\| = |c|\|\vec{x}\|$ for all scalars $c \in \mathbb{R}$ and vectors $\vec{x} \in \mathbb{R}^n$ ("absolute scalability")
- $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ ("triangle inequality")

Other than $\|\cdot\|_2$, there are many examples of norms:

- The p-norm $\|\vec{x}\|_p$, for $p \geq 1$, given by:

$$\|\vec{x}\|_p \equiv \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}$$

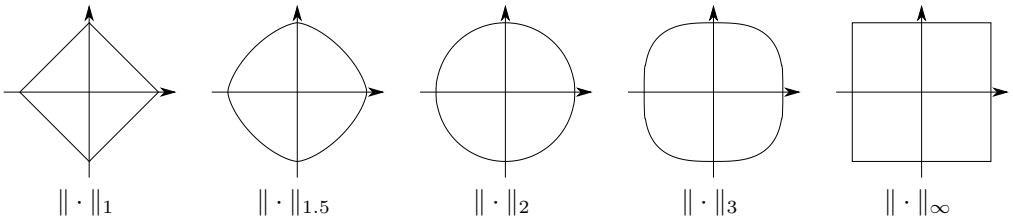


FIGURE 3.7 The set $\{\vec{x} \in \mathbb{R}^2 : \|\vec{x}\| = 1\}$ for different vector norms $\|\cdot\|$.

Of particular importance is the 1-norm, also known as the "Manhattan" or "taxicab" norm, given by

$$\|\vec{x}\|_1 \equiv \sum_{k=1}^{n} |x_k|.$$

This norm receives its nickname because it represents the distance a taxicab drives between two points in a city where the roads only run north/south and east/west.

- The ∞-norm $\|\vec{x}\|_\infty$ given by:

$$\|\vec{x}\|_\infty \equiv \max(|x_1|, |x_2|, \cdots, |x_n|).$$

These norms are illustrated in Figure 3.7 by showing the "unit circle" $\{\vec{x} \in \mathbb{R}^2 : \|\vec{x}\| = 1\}$; this visualization shows that $\|\vec{v}\|_p \leq \|\vec{v}\|_q$ when $p > q$.

Despite these geometric differences, many norms on $\mathbb{R}^n$ have similar behavior. In particular, suppose we say two norms are *equivalent* when they satisfy the following property:

Definition 3.3 (Equivalent norms)**.** Two norms $\|\cdot\|$ and $\|\cdot\|'$ are *equivalent* if there exist constants c_{low} and c_{high} such that $c_{\text{low}}\|\vec{x}\| \leq \|\vec{x}\|' \leq c_{\text{high}}\|\vec{x}\|$ for all $\vec{x} \in \mathbb{R}^n$.

This condition guarantees that up to some constant factors, all norms agree on which vectors are "small" and "large." We will state without proof a famous theorem from analysis:

Theorem 3.2 (Equivalence of norms on $\mathbb{R}^n$)**.** All norms on $\mathbb{R}^n$ are equivalent.

This somewhat surprising result implies that all vector norms have the same *rough* behavior, but the choice of a norm for analyzing or stating a particular problem can make a huge difference. For instance, on $\mathbb{R}^3$ the ∞-norm considers the vector $(1000, 1000, 1000)$ to have the same norm as $(1000, 0, 0)$ whereas the 2-norm certainly is affected by the additional nonzero values.

Since we perturb not only vectors but also matrices, we must also be able to take the norm of a matrix. The definition of a matrix norm is nothing more than Definition 3.2 with matrices in place of vectors. For this reason, we can "unroll" any matrix in $\mathbb{R}^{m\times n}$ to a vector in $\mathbb{R}^{nm}$ to adapt any vector norm to matrices. One such norm is the *Frobenius norm*, given by

$$\|A\|_{\text{Fro}} \equiv \sqrt{\sum_{i,j} a_{ij}^2}.$$

Such adaptations of vector norms, however, are not always meaningful. In particular, norms on matrices A constructed this way may not have a clear connection to the *action* of A on vectors. Since we usually use matrices to encode linear transformations, we would

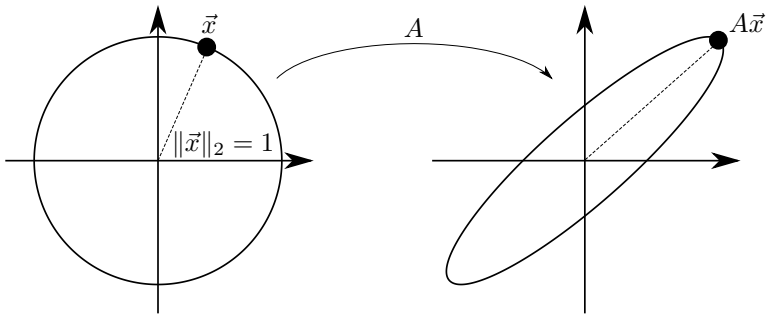


FIGURE 3.8 The norm $\|\cdot\|_2$ induces a matrix norm measuring the largest distortion of any point the unit circle after applying A.

prefer a norm that helps us understand what happens when A is multiplied by different vectors $\vec{x}$. With this motivation, we can define the matrix norm *induced* by a vector norm as follows:

Definition 3.4 (Induced norm)**.** The matrix norm on $\mathbb{R}^{m\times n}$ *induced* by a norm $\|\cdot\|$ on $\mathbb{R}^n$ is given by

$$\|A\| \equiv \max\{\|A\vec{x}\| : \|\vec{x}\| = 1\}.$$

That is, the induced norm is the maximum length of the image of a unit vector multiplied by A.

This definition in the case $\|\cdot\| = \|\cdot\|_2$ is illustrated in Figure 3.8. Since vector norms satisfy $\|c\vec{x}\| = |c|\|\vec{x}\|$, this definition is equivalent to requiring

$$\|A\| \equiv \max_{\vec{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|A\vec{x}\|}{\|\vec{x}\|}.$$

From this standpoint, the norm of A induced by $\|\cdot\|$ is the largest achievable ratio of the norm of $A\vec{x}$ relative to that of the input $\vec{x}$.

This general definition makes it somewhat complicated to compute the norm $\|A\|$ given a matrix A and a choice of $\|\cdot\|$. Fortunately, the matrix norms induced by many popular vector norms can be simplified. Some well-known formulae for matrix norms include the following:

- The induced one-norm of A is the maximum absolute column sum of A:

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$$

- The induced ∞-norm of A is the maximum absolute row sum of A:

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$$

- The induced two-norm, or *spectral norm*, of $A \in \mathbb{R}^{n\times n}$ is the square root of the largest eigenvalue of $A^\top A$. That is,

$$\|A\|_2^2 = \max\{\lambda : \text{there exists } \vec{x} \in \mathbb{R}^n \text{ with } A^\top A\vec{x} = \lambda\vec{x}\}$$

The first two norms are computable directly from the elements of A; the third will require machinery from Chapter 6.

3.3.2 Condition Numbers

Now that we have tools for measuring the action of a matrix, we can define the *condition number* of a linear system by adapting our generic definition of condition numbers from Chapter 1. In this section, we will follow the development presented in [46].

Suppose we are given a perturbation δA of a matrix A and a perturbation $\delta\vec{b}$ of the right-hand side of the linear system $A\vec{x} = \vec{b}$. For small values of ε, ignoring invertibility technicalities we can write a vector-valued function $\vec{x}(\varepsilon)$ as the solution to

$$(A + \varepsilon \cdot \delta A)\vec{x}(\varepsilon) = \vec{b} + \varepsilon \cdot \delta\vec{b}.$$

Differentiating both sides with respect to ε and applying the product rule shows:

$$\delta A \cdot \vec{x}(\varepsilon) + (A + \varepsilon \cdot \delta A)\frac{d\vec{x}(\varepsilon)}{d\varepsilon} = \delta\vec{b}.$$

In particular, when $\varepsilon = 0$ we find

$$\delta A \cdot \vec{x}(0) + A\frac{d\vec{x}}{d\varepsilon}\Big|_{\varepsilon=0} = \delta\vec{b}$$

or, equivalently,

$$\frac{d\vec{x}}{d\varepsilon}\Big|_{\varepsilon=0} = A^{-1}(\delta\vec{b} - \delta A \cdot \vec{x}(0)).$$

Using the Taylor expansion, we can write

$$\vec{x}(\varepsilon) = \vec{x} + \varepsilon\vec{x}'(0) + O(\varepsilon^2),$$

where we define $\vec{x}'(0) = \frac{d\vec{x}}{d\varepsilon}\big|_{\varepsilon=0}$. Thus, we can expand the relative error made by solving the perturbed system:

$$\begin{aligned}
\frac{\|\vec{x}(\varepsilon) - \vec{x}(0)\|}{\|\vec{x}(0)\|} &= \frac{\|\varepsilon\vec{x}'(0) + O(\varepsilon^2)\|}{\|\vec{x}(0)\|} \text{ by the Taylor expansion above} \\
&= \frac{\|\varepsilon A^{-1}(\delta\vec{b} - \delta A \cdot \vec{x}(0)) + O(\varepsilon^2)\|}{\|\vec{x}(0)\|} \text{ by the derivative we computed} \\
&\leq \frac{|\varepsilon|}{\|\vec{x}(0)\|}(\|A^{-1}\delta\vec{b}\| + \|A^{-1}\delta A \cdot \vec{x}(0))\|) + O(\varepsilon^2) \\
&\qquad \text{by the triangle inequality } \|A + B\| \leq \|A\| + \|B\| \\
&\leq |\varepsilon|\|A^{-1}\|\left(\frac{\|\delta\vec{b}\|}{\|\vec{x}(0)\|} + \|\delta A\|\right) + O(\varepsilon^2) \text{ by the identity } \|AB\| \leq \|A\|\|B\| \\
&= |\varepsilon|\|A^{-1}\|\|A\|\left(\frac{\|\delta\vec{b}\|}{\|A\|\|\vec{x}(0)\|} + \frac{\|\delta A\|}{\|A\|}\right) + O(\varepsilon^2) \\
&\leq |\varepsilon|\|A^{-1}\|\|A\|\left(\frac{\|\delta\vec{b}\|}{\|A\vec{x}(0)\|} + \frac{\|\delta A\|}{\|A\|}\right) + O(\varepsilon^2) \text{ since } \|A\vec{x}(0)\| \leq \|A\|\|\vec{x}(0)\| \\
&= |\varepsilon|\|A^{-1}\|\|A\|\left(\frac{\|\delta\vec{b}\|}{\|\vec{b}\|} + \frac{\|\delta A\|}{\|A\|}\right) + O(\varepsilon^2) \text{ since by definition } A\vec{x}(0) = \vec{b}.
\end{aligned}$$

Here we have applied some properties of the matrix norm which follow from corresponding properties for vectors; you will check them explicitly in Problem 3.10.

The sum $D \equiv \|\delta\vec{b}\|/\|\vec{b}\| + \|\delta A\|/\|A\|$ appearing in the last equality above encodes the magnitudes of the perturbations of δA and $\delta\vec{b}$ relative to the magnitudes of A and $\vec{b}$, resp. From this standpoint, to first order we have bounded the relative error of perturbing the system by ε in terms of the factor $\kappa \equiv \|A\|\|A^{-1}\|$:

$$\frac{\|\vec{x}(\varepsilon) - \vec{x}(0)\|}{\|\vec{x}(0)\|} \leq \varepsilon \cdot D \cdot \kappa + O(\varepsilon^2)$$

Hence, the quantity κ bounds the conditioning of linear systems involving A, inspiring the following definition:

Definition 3.5 (Matrix condition number)**.** The condition number of $A \in \mathbb{R}^{n\times n}$ for a given matrix norm $\|\cdot\|$ is

$$\text{cond } A \equiv \|A\|\|A^{-1}\|.$$

If A is not invertible, we take cond $A \equiv \infty$.

For nearly any matrix norm, cond $A \geq 1$ for all A. Scaling A has no effect on its condition number; generally, the condition number of the identity matrix is 1. Large condition numbers indicate that solutions to $A\vec{x} = \vec{b}$ are unstable under perturbations of A or $\vec{b}$.

If $\|\cdot\|$ is induced by a vector norm and A is invertible, then we have

$$\begin{aligned}
\|A^{-1}\| &= \max_{\vec{x}\neq\vec{0}} \frac{\|A^{-1}\vec{x}\|}{\|\vec{x}\|} \text{ by definition} \\
&= \max_{\vec{y}\neq\vec{0}} \frac{\|\vec{y}\|}{\|A\vec{y}\|} \text{ by substituting } \vec{y} = A^{-1}\vec{x} \\
&= \left(\min_{\vec{y}\neq\vec{0}} \frac{\|A\vec{y}\|}{\|\vec{y}\|}\right)^{-1} \text{ by taking the reciprocal}
\end{aligned}$$

In this case, the condition number of A is given by:

$$\text{cond } A = \left(\max_{\vec{x}\neq\vec{0}} \frac{\|A\vec{x}\|}{\|\vec{x}\|}\right)\left(\min_{\vec{y}\neq\vec{0}} \frac{\|A\vec{y}\|}{\|\vec{y}\|}\right)^{-1}$$

In other words, cond A measures the ratio of the maximum to the minimum possible stretch of a vector $\vec{x}$ under A; this interpretation is illustrated in Figure 3.9.

A desirable stability property of a system $A\vec{x} = \vec{b}$ is that if A or $\vec{b}$ is perturbed, the solution $\vec{x}$ does not change considerably. Our motivation for cond A shows that when the condition number is small, the change in $\vec{x}$ is small relative to the change in A or $\vec{b}$. Otherwise, a small change in the parameters of the linear system can cause large deviations in $\vec{x}$; this instability can cause linear solvers to make large mistakes in $\vec{x}$ due to rounding and other approximations during the solution process.

In practice, we might wish to evaluate cond A before solving $A\vec{x} = \vec{b}$ to see how successful we can expect to be in this process. Taking the norm $\|A^{-1}\|$, however, can be as difficult as computing the full inverse A^{-1}. A subtle "chicken-and-egg problem" exists here: Do we need to compute the condition number of computing matrix condition numbers? A common way out is to *bound* or *approximate* cond A using expressions that are easier to evaluate. Lower bounds on the condition number represent optimistic bounds that can be used to cull out particularly bad matrices A, while upper bounds guarantee behavior in the worst case.

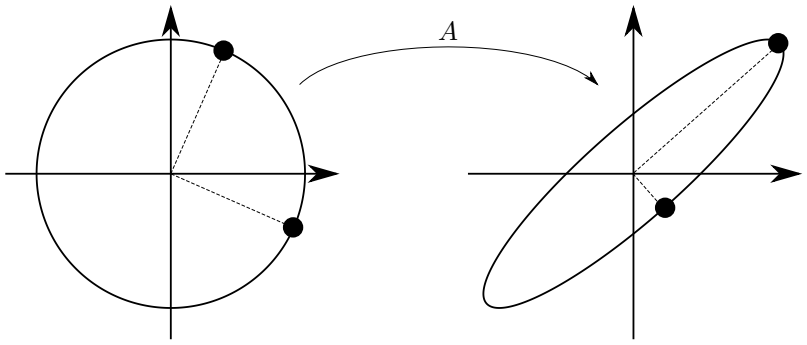


FIGURE 3.9 The condition number of A measures the ratio of the largest to smallest distortion of any point on the unit circle mapped under A.

The problem of condition number estimation is itself an area of active research in numerical analysis.

For example, one way to lower-bound the condition number is to apply the identity $\|A^{-1}\vec{x}\| \leq \|A^{-1}\|\|\vec{x}\|$ as in Problem 3.10. Then, for any $\vec{x} \neq \vec{0}$ we can write $\|A^{-1}\| \geq {}^{\|A^{-1}\vec{x}\|}/_{\|\vec{x}\|}$. Thus,

$$\text{cond } A = \|A\|\|A^{-1}\| \geq \frac{\|A\|\|A^{-1}\vec{x}\|}{\|\vec{x}\|}.$$

So, we can bound the condition number by solving $A^{-1}\vec{x}$ for some vectors $\vec{x}$. The necessity of a linear solver to find $A^{-1}\vec{x}$ again creates a circular dependence on the condition number to evaluate the quality of the estimate! After considering eigenvalue problems, in future chapters we will provide more reliable estimates when $\|\cdot\|$ is induced by the two-norm.

3.4 EXERCISES

3.1 ("Matrix calculus") The optimization problem we posed for $A \in \mathbb{R}^{2\times 2}$ in §3.1.4 is an example of a problem where the unknown is a matrix rather than a vector. These problems appear frequently in machine learning and have inspired an alternative notation for differential calculus better suited to calculations of this sort.

(a) Suppose $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a smooth function. Justify why the gradient of f can be thought of as an $n \times m$ matrix. We will use the notation $\frac{\partial f}{\partial A}$ to notate the gradient of $f(A)$ with respect to A.

(b) Take the gradient ${}^{\partial}/_{\partial A}$ of the following functions, assuming $\vec{x}$ and $\vec{y}$ are constant vectors:
   (i) $\vec{x}^\top A\vec{y}$
   (ii) $\vec{x}^\top A^\top A\vec{x}$
   (iii) $(\vec{x} - A\vec{y})^\top W(\vec{x} - A\vec{y})$ for symmetric W

(c) Now, suppose $X \in \mathbb{R}^{m\times n}$ is a smooth function of a scalar variable $X(t) : \mathbb{R} \to \mathbb{R}^{m\times n}$. We can notate the *differential* $\partial X \equiv X'(t)$. For matrix functions $X(t)$ and $Y(t)$, justify the following identities:
   (i) $\partial(X + Y) = \partial X + \partial Y$
   (ii) $\partial(X^\top) = (\partial X)^\top$

(iii) $\partial(XY) = (\partial X)Y + X(\partial Y)$
(iv) $\partial(X^{-1}) = -X^{-1}(\partial X)X^{-1}$ (see problem 0.11)

After establishing a dictionary of identities like the ones above, taking the derivatives of functions involving matrices becomes a far less cumbersome task. See [94] for a comprehensive reference of identities and formulas in matrix calculus.

3.2 The system of equations for A and $\vec{b}$ in §3.1.4 must be "unrolled" if we wish to use standard software for solving linear systems of equations to recover the image transformation. Define

$$A \equiv \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad \text{and} \qquad \vec{b} \equiv \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

We can combine all our unknowns into a vector $\vec{u}$ as follows:

$$\vec{u} \equiv \begin{pmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \\ b_1 \\ b_2 \end{pmatrix}.$$

Write a matrix $M \in \mathbb{R}^{6\times 6}$ and vector $\vec{d} \in \mathbb{R}^6$ so that $\vec{u}$—and hence A and $\vec{b}$—can be recovered by solving the system $M\vec{u} = \vec{d}$ for $\vec{u}$; you can use any computable temporary variables to simplify your notation, including $\vec{x}_{\text{sum}}$, $\vec{y}_{\text{sum}}$, X, and C.

3.3 There are many ways to motivate the harmonic parameterization technique from §3.1.6. One alternative is to consider the *Dirichlet energy* of a parameterization

$$E_D[\vec{t}(\cdot)] \equiv \sum_{(v,w)\in E} \|\vec{t}(v) - \vec{t}(w)\|_2^2.$$

Then, we can write an optimization problem given boundary vertex positions $\vec{t}_0(\cdot) : B \to \mathbb{R}^2$:

$$\begin{array}{ll} \text{minimize} & E_D[\vec{t}(\cdot)] \\ \text{such that} & \vec{t}(v) = \vec{t}_0(v)\ \forall v \in B. \end{array}$$

This optimization minimizes the Dirichlet energy $E_D[\cdot]$ over all possible parameterizations $\vec{t}(\cdot)$ with the constraint that the positions of boundary vertices $v \in B$ are fixed. Show that after minimizing this energy, interior vertices $v \in V\backslash B$ satisfy the barycenter property introduced in §3.1.6:

$$\vec{t}(v) = \frac{1}{|N(v)|} \sum_{w\in N(v)} \vec{t}(w).$$

This variational formulation connects the technique to the differential geometry of smooth maps into the plane.

3.4 A more general version of the Cholesky decomposition that does not require the computation of square roots is the LDLT decomposition.

(a) Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and admits an LU factorization (without pivoting). Show that A can be factored $A = LDL^\top$, where D is diagonal and L is lower-triangular.
Hint: Take $D \equiv UL^{-\top}$; you must show that this matrix is diagonal.

(b) Modify the construction of the Cholesky decomposition from §3.2.1 to show how a symmetric, positive-definite matrix A can be factored $A = LDL^\top$ without using any square root operations. Does your algorithm only work when A is positive definite?

3.5 Suppose $A \in \mathbb{R}^{m\times n}$ has full rank, where $m < n$. Show that taking $\vec{x} = A^\top(AA^\top)^{-1}\vec{b}$ solves the following optimization problem:

$$\begin{array}{rl} \min_{\vec{x}} & \|\vec{x}\|_2 \\ \text{such that} & A\vec{x} = \vec{b}. \end{array}$$

Furthermore, show that taking $\alpha \to 0$ in the Tikhonov-regularized system from §3.1.3 recovers this choice of $\vec{x}$.

3.6 Suppose $A \in \mathbb{R}^{n\times n}$ is *tridiagonal*, meaning it can be written:

$$A = \begin{pmatrix} v_1 & w_1 & & & & & \\ u_2 & v_2 & w_2 & & & & \\ & u_3 & v_3 & w_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & u_{n-1} & v_{n-1} & w_{n-1} \\ & & & & u_n & v_n \end{pmatrix}.$$

Show that in this case the system $A\vec{x} = \vec{b}$ can be solved in $O(n)$ time. You can assume that A is *diagonally dominant*, meaning $|v_i| > |u_i| + |w_i|$ for all i.
Hint: Start from Gaussian elimination. This algorithm usually is attributed to [112].

3.7 Show how linear techniques can be used to solve the following optimization problem for $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{k\times n}, \vec{c} \in \mathbb{R}^k$:

$$\begin{array}{rl} \text{minimize}_{\vec{x}\in\mathbb{R}^n} & \|A\vec{x}\|_2^2 \\ \text{such that} & B\vec{x} = \vec{c} \end{array}$$

3.8 Suppose $A \in \mathbb{R}^{n\times n}$ admits a Cholesky factorization $A = LL^\top$.

(a) Show that A must be positive semidefinite.

(b) Use this observation to suggest an algorithm for checking if a matrix is positive semidefinite.

3.9 Are all matrix norms on $\mathbb{R}^{m\times n}$ equivalent? Why or why not?

3.10 For this problem, assume that the matrix norm $\|A\|$ for $A \in \mathbb{R}^{n\times n}$ is induced by a vector norm $\|\vec{v}\|$ for $\vec{v} \in \mathbb{R}^n$ (but it may be the case that $\|\cdot\| \neq \|\cdot\|_2$).

(a) For $A, B \in \mathbb{R}^{n\times n}$, show $\|A + B\| \leq \|A\| + \|B\|$.

(b) For $A, B \in \mathbb{R}^{n\times n}$ and $\vec{v} \in \mathbb{R}^n$, show $\|A\vec{v}\| \leq \|A\|\|\vec{v}\|$ and $\|AB\| \leq \|A\|\|B\|$.

(c) For $k > 0$ and $A \in \mathbb{R}^{n\times n}$, show $\|A^k\|^{1/k} \geq |\lambda|$ for any real eigenvalue λ of A.

(d) For $A \in \mathbb{R}^{n\times n}$ and $\|\vec{v}\|_1 \equiv \sum_i |v_i|$, show $\|A\|_1 = \max_j \sum_i |a_{ij}|$.

(e) Prove Gelfand's formula: $\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$, where $\rho(A) \equiv \max\{|\lambda_i|\}$ for eigenvalues $\lambda_1, \ldots, \lambda_m$ of A. In fact, this formula holds for any matrix norm $\|\cdot\|$.

3.11 ("Screened Poisson smoothing") Suppose we sample a function $f(x)$ at n positions $x_1, x_2, \ldots, x_n$, yielding a point $\vec{y} \equiv (f(x_1), f(x_2), \ldots, f(x_n)) \in \mathbb{R}^n$. Our measurements might be noisy, however, so a common task in graphics and statistics is to smooth these values to obtain a new vector $\vec{z} \in \mathbb{R}^n$.

(a) Provide least-squares energy terms measuring the following:
   (i) The similarity of $\vec{y}$ and $\vec{z}$.
   (ii) The smoothness of $\vec{z}$.
   Hint: We expect $f(x_{i+1}) - f(x_i)$ to be small for smooth f.

(b) Propose an optimization problem for smoothing $\vec{y}$ using the terms above to obtain $\vec{z}$, and argue that it can be solved using linear techniques.

(c) Suppose n is very large. What properties of the matrix in 3.11b might be relevant in choosing an effective algorithm to solve the linear system?

3.12 ("Kernel trick") In this chapter, we covered techniques for linear and nonlinear *parametric* regression. Now, we will develop one least-squares technique for *nonparametic* regression that is used commonly in machine learning and vision.

(a) You can think of the least-squares problem as learning the vector $\vec{a}$ in a function $f(\vec{x}) = \vec{a} \cdot \vec{x}$ given a number of examples $\vec{x}^{(1)} \mapsto y^{(1)}, \ldots, \vec{x}^{(k)} \mapsto y^{(k)}$ and the assumption $f(\vec{x}^{(i)}) \approx y^{(i)}$. Suppose the columns of X are the vectors $\vec{x}^{(i)}$ and that $\vec{y}$ is the vector of values $y^{(i)}$. Provide the normal equations for recovering $\vec{a}$ with Tikhonov regularization.

(b) Show that $\vec{a} \in \text{span}\,\{\vec{x}^{(1)}, \ldots, \vec{x}^{(k)}\}$ in the Tikhonov-regularized system.

(c) Thus, we can write $\vec{a} = c_1\vec{x}^{(1)} + \cdots + c_k\vec{x}^{(k)}$. Give a $k \times k$ linear system of equations satisfied by $\vec{c}$ assuming $X^\top X$ is invertible.

(d) One way to do nonlinear regression might be to write a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$ and learn $f_\phi(\vec{x}) = \vec{a} \cdot \phi(\vec{x})$, where ϕ may be nonlinear. Define $K(\vec{x}, \vec{y}) = \phi(\vec{x}) \cdot \phi(\vec{y})$. Assuming we continue to use regularized least squares as in 3.12a, give an alternative form of f_ϕ that can be computed by evaluating K rather than ϕ. (Hint: What are the elements of $X^\top X$?)

(e) Consider the following formula from the Fourier transform of the Gaussian:

$$e^{-\pi(s-t)^2} = \int_{-\infty}^{\infty} e^{-\pi x^2}(\sin(2\pi sx)\sin(2\pi tx) + \cos(2\pi sx)\cos(2\pi tx))\,dx$$

Suppose we wrote $K(x, y) = e^{-\pi(x-y)^2}$. Explain how this "looks like" $\phi(x) \cdot \phi(y)$ for some ϕ. How does this suggest that the technique from 3.12d can be generalized?

3.13 ("Discrete Fourier transform") This problem deals with complex numbers, so we will take $i \equiv \sqrt{-1}$.

(a) Suppose $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$. Derive *de Moivre's formula* by induction on n:

$$(\cos\theta + i\sin\theta)^n = \cos n\theta + i\sin n\theta.$$

(b) Euler's formula uses "complex exponentials" to define $e^{i\theta} \equiv \cos\theta + i\sin\theta$. Write de Moivre's formula in this notation.

(c) Define the *primitive n-th root of unity* as $\omega_n \equiv e^{-2\pi i/n}$. The *discrete Fourier transform* matrix can be written

$$W_n \equiv \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \omega_n^3 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \omega_n^6 & \cdots & \omega_n^{2(n-1)} \\ 1 & \omega_n^3 & \omega_n^6 & \omega_n^9 & \cdots & \omega_n^{3(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \omega_n^{3(n-1)} & \cdots & \omega_n^{(n-1)(n-1)} \end{pmatrix}.$$

Show that W_n can be written in terms of a Vandermonde matrix, as defined in Example 3.3.

(d) The *complex conjugate* of $a + bi \in \mathbb{C}$, where $a, b \in \mathbb{R}$, is $\overline{a+bi} \equiv a - bi$. Show that $W_n^{-1} = W_n^*$, where $W_n^* \equiv \overline{W}^\top$.

(e) Suppose $n = 2^k$. In this case, show how W_n can be applied to a vector $\vec{x} \in \mathbb{C}^n$ via two applications of $W_{n/2}$ and post-processing that takes $O(n)$ time.

Note: The *fast Fourier transform* essentially uses this technique recursively to apply W_n in $O(n \log n)$ time.

CHAPTER 4

Column Spaces and QR

CONTENTS

4.1 The Structure of the Normal Equations 87
4.2 Orthogonality 88
4.3 Strategy for Non-Orthogonal Matrices 89
4.4 Gram-Schmidt Orthogonalization 90
4.4.1 Projections 90
4.4.2 Gram-Schmidt Algorithm 92
4.5 Householder Transformations 94
4.6 Reduced QR Factorization 96

ONE way to interpret the linear problem $A\vec{x} = \vec{b}$ for $\vec{x}$ is that we wish to write $\vec{b}$ as a linear combination of the columns of A with weights given in $\vec{x}$. This perspective does not change when we allow $A \in \mathbb{R}^{m\times n}$ to be non-square, but the solution may not exist or be unique depending on the structure of the column space of A. For these reasons, some techniques for factoring matrices and analyzing linear systems seek simpler *representations* of the column space to disambiguate solvability and span more explicitly than row-based factorizations like LU.

4.1 THE STRUCTURE OF THE NORMAL EQUATIONS

As showed in §3.1.2, a necessary and sufficient condition for $\vec{x}$ to be a solution of the least-squares problem $A\vec{x} \approx \vec{b}$ is that $\vec{x}$ must satisfy the normal equations $(A^\top A)\vec{x} = A^\top\vec{b}$. This theorem suggests least-squares problems can be solved using linear techniques on the matrix $A^\top A$. With this observation in mind, methods such as Cholesky factorization employ the special structure of least-squares problems to the solver's advantage.

There is one large problem limiting the use of the normal equations, however. For now, suppose A is square; then we can write:

$$\begin{aligned}\text{cond } A^\top A &= \|A^\top A\|\|(A^\top A)^{-1}\| \\ &\approx \|A^\top\|\|A\|\|A^{-1}\|\|(A^\top)^{-1}\| \text{ for many choices of } \|\cdot\| \\ &= \|A\|^2\|A^{-1}\|^2 \\ &= (\text{cond } A)^2\end{aligned}$$

That is, the condition number of $A^\top A$ is approximately the **square** of the condition number of A! Thus, while generic linear strategies might work on $A^\top A$ when the least-squares problem is "easy," when the columns of A are nearly linearly dependent these strategies are likely to generate considerable error since they do not deal with A directly.

Intuitively, a primary reason that cond $A^\top A$ can be large illustrated in Figure 4.1 is

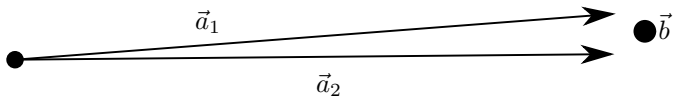


FIGURE 4.1 The vectors $\vec{a}_1$ and $\vec{a}_2$ nearly coincide; hence, writing $\vec{b}$ in the span of these vectors is difficult since $\vec{v}_1$ can be replaced with $\vec{v}_2$ or vice versa in a linear combination without incurring much error.

that columns of A might look "similar." Think of each column of A as a vector in $\mathbb{R}^m$. If two columns $\vec{a}_i$ and $\vec{a}_j$ satisfy $\vec{a}_i \approx \vec{a}_j$, then the least-squares residual length $\|\vec{b} - A\vec{x}\|_2$ will not suffer much if we replace multiples of $\vec{a}_i$ with multiples of $\vec{a}_j$ or vice versa. This wide range of nearly—but not completely—equivalent solutions yields poor conditioning. While the resulting vector $\vec{x}$ is unstable, however, the product $A\vec{x}$ remains nearly unchanged, by design of our substitution; if our goal is to write $\vec{b}$ in the column space of A, either solution would suffice.

To solve such poorly-conditioned problems, we will employ an alternative technique with closer attention to the column space of A rather than employing row operations as in Gaussian elimination. This strategy identifies and deals with such near-dependencies *explicitly*, bringing about greater numerical stability.

4.2 ORTHOGONALITY

We have determined when the least-squares problem is difficult, but we might also ask when it is most straightforward. If we can reduce a system to the straightforward case without inducing conditioning problems along the way, we will have found a stable way around the issues explained in §4.1.

The easiest linear system to solve is $I_{n\times n}\vec{x} = \vec{b}$, where $I_{n\times n}$ is the $n \times n$ identity matrix: The solution is $\vec{x} \equiv \vec{b}$! We are unlikely to bother using a linear solver to invert this particular linear system on purpose, but we may do so accidentally while solving least-squares. Even when $A \neq I_{n\times n}$—A need not be square—we may in particularly lucky circumstances find that the normal matrix $A^\top A$ satisfies $A^\top A = I_{n\times n}$, making the least-squares solve trivial. To avoid confusion with the general case, we will use the letter Q to represent such a matrix satisfying $Q^\top Q = I_{n\times n}$.

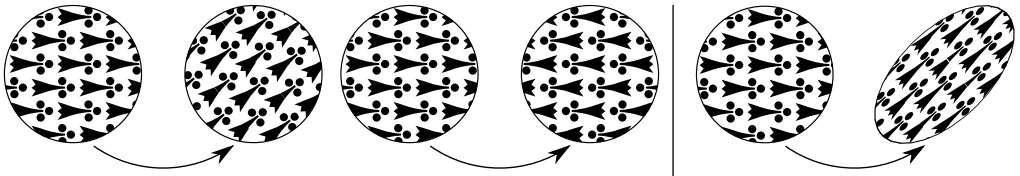
Praying that $Q^\top Q = I_{n\times n}$ unlikely will yield a useful algorithm, but we can examine this case to see how it becomes so favorable. Write the columns of Q as vectors $\vec{q}_1, \cdots, \vec{q}_n \in \mathbb{R}^m$. Then, the product $Q^\top Q$ has the following structure:

$$Q^\top Q = \begin{pmatrix} - & \vec{q}_1^\top & - \\ - & \vec{q}_2^\top & - \\ & \vdots & \\ - & \vec{q}_n^\top & \end{pmatrix} \begin{pmatrix} | & | & & | \\ \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} \vec{q}_1 \cdot \vec{q}_1 & \vec{q}_1 \cdot \vec{q}_2 & \cdots & \vec{q}_1 \cdot \vec{q}_n \\ \vec{q}_2 \cdot \vec{q}_1 & \vec{q}_2 \cdot \vec{q}_2 & \cdots & \vec{q}_2 \cdot \vec{q}_n \\ \vdots & \vdots & \cdots & \vdots \\ \vec{q}_n \cdot \vec{q}_1 & \vec{q}_n \cdot \vec{q}_2 & \cdots & \vec{q}_n \cdot \vec{q}_n \end{pmatrix}$$

Setting the expression on the right equal to $I_{n\times n}$ yields the following relationship:

$$\vec{q}_i \cdot \vec{q}_j = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j \end{cases}$$

In other words, the columns of Q are unit-length and orthogonal to one another. We say that they form an *orthonormal basis* for the column space of Q:



(a) Isometric (b) Not isometric

FIGURE 4.2 Isometries can rotate and flip vectors (a) but cannot stretch or shear them (b).

Definition 4.1 (Orthonormal; orthogonal matrix)**.** A set of vectors $\{\vec{v}_1, \cdots, \vec{v}_k\}$ is *orthonormal* if $\|\vec{v}_i\| = 1$ for all i and $\vec{v}_i \cdot \vec{v}_j = 0$ for all $i \neq j$. A square matrix whose columns are orthonormal is called an *orthogonal* matrix.

The standard basis $\{\vec{e}_1, \vec{e}_2, \ldots, \vec{e}_n\}$ is a typical example of an orthonormal basis, and since the columns of the identity matrix $I_{n\times n}$ are these vectors we know $I_{n\times n}$ is an orthogonal matrix.

We motivated our discussion by asking when we can expect $Q^\top Q = I_{n\times n}$. Now we know that this condition occurs exactly when the columns of Q are orthonormal. Furthermore, if Q is square and invertible with $Q^\top Q = I_{n\times n}$, then by multiplying both sides of the expression $Q^\top Q = I_{n\times n}$ by Q^{-1} shows $Q^{-1} = Q^\top$. Hence, $Q\vec{x} = \vec{b}$ is equivalent to $\vec{x} = Q^\top \vec{b}$ after multiplying both sides by the transpose $Q^\top$.

Orthonormality has a strong geometric interpretation. Recall from Chapter 0 that we can regard two orthogonal vectors $\vec{a}$ and $\vec{b}$ as being *perpendicular*. So, an orthonormal set of vectors is a set of mutually-perpendicular unit vectors in $\mathbb{R}^n$. Furthermore, if Q is orthogonal, then its action does not affect the length of vectors:

$$\|Q\vec{x}\|^2 = \vec{x}^\top Q^\top Q\vec{x} = \vec{x}^\top I_{n\times n}\vec{x} = \vec{x} \cdot \vec{x} = \|\vec{x}\|^2$$

Similarly, Q cannot affect the angle between two vectors, since:

$$(Q\vec{x}) \cdot (Q\vec{y}) = \vec{x}^\top Q^\top Q\vec{y} = \vec{x}^\top I_{n\times n}\vec{y} = \vec{x} \cdot \vec{y}$$

From this standpoint, if Q is orthogonal, then the operation $\vec{x} \mapsto Q\vec{x}$ is an *isometry* of $\mathbb{R}^n$, that is, it preserves lengths and angles. In other words, as illustrated in Figure 4.2 it can rotate or reflect vectors but cannot scale or shear them. From a high level, the linear algebra of orthogonal matrices is easier because their actions do not affect the geometry of the underlying space.

4.3 STRATEGY FOR NON-ORTHOGONAL MATRICES

Most matrices encountered when solving $A\vec{x} = \vec{b}$ or the least-squares problem $A\vec{x} \approx \vec{b}$ will not be orthogonal, so the machinery of §4.2 does not apply directly. For this reason, we must do some additional computations to connect the general case to the orthogonal one. In the end, we will derive an alternative to LU factorization using orthogonal rather than substitution matrices.

Take a matrix $A \in \mathbb{R}^{m\times n}$, and denote its column space as col A; col A is the span of the columns of A. Now, suppose a matrix $B \in \mathbb{R}^{n\times n}$ is invertible. We make the following observation about the column space of AB relative to that of A:

Proposition 4.1 (Column space invariance)**.** For any $A \in \mathbb{R}^{m\times n}$ and invertible $B \in \mathbb{R}^{n\times n}$, col $A =$ col AB.

Proof. Suppose $\vec{b} \in$ col A. Then, by definition there exists $\vec{x}$ with $A\vec{x} = \vec{b}$. Then, if we take $\vec{y} = B^{-1}\vec{x}$, then $AB\vec{y} = (AB)\cdot(B^{-1}\vec{x}) = A\vec{x} = \vec{b}$, so $\vec{b} \in$ col AB. Conversely, take $\vec{c} \in$ col AB, so there exists $\vec{y}$ with $(AB)\vec{y} = \vec{c}$. Then, $A\cdot(B\vec{y}) = \vec{c}$, showing that $\vec{c}$ is in col A. □

Recall the "elimination matrix" description of Gaussian elimination: We started with a matrix A and applied row operation matrices E_i such that the sequence $A, E_1A, E_2E_1A, \ldots$ eventually reduced to more easily-solved triangular systems. The proposition above suggests an alternative strategy for situations like least-squares in which we care about the column space of A: Apply *column* operations to A by *post*-multiplication until the columns are orthonormal. So long as these operations are invertible, the Proposition 4.1 shows that the column space of the modified matrices will be the same as that of A.

In the end, we will attempt to find a product $Q = AE_1E_2\cdots E_k$ starting from A and applying invertible operation matrices E_i such that Q is orthonormal. As we have argued above, the proposition shows that col $Q =$ col A. Inverting these operations yields a factorization $A = QR$ for $R = E_k^{-1}E_{k-1}^{-1}\cdots E_1^{-1}$. The columns of the matrix Q contain an orthonormal basis for the column space of A, and with careful design we can once again make R upper triangular.

In this case, when $A = QR$, by orthogonality of Q we have $A^\top A = R^\top Q^\top QR = R^\top R$. Thus, the normal equations $A^\top A\vec{x} = A^\top\vec{b}$ imply $R^\top R\vec{x} = R^\top Q^\top\vec{b}$, or equivalently $R\vec{x} = Q^\top\vec{b}$. If we design R to be a triangular matrix, then solving the least-squares system $A^\top A\vec{x} = A^\top\vec{b}$ can be carried out efficiently by back-substitution via $R\vec{x} = Q^\top\vec{b}$.

4.4 GRAM-SCHMIDT ORTHOGONALIZATION

Our first algorithm for finding QR factorizations follows naturally from our discussion above but may suffer from numerical issues. We use it here as an initial example of orthogonalization and then will improve upon it with better operations.

4.4.1 Projections

Suppose we have two vectors $\vec{a}$ and $\vec{b}$, with $\vec{a} \neq \vec{0}$. Then, we could easily ask, "Which multiple of $\vec{a}$ is closest to $\vec{b}$?" Mathematically, this task is equivalent to minimizing $\|c\vec{a} - \vec{b}\|_2^2$ over all possible $c \in \mathbb{R}$. If we think of $\vec{a}$ and $\vec{b}$ as $n \times 1$ matrices and c as a 1×1 matrix, then this is nothing more than an unconventional least-squares problem $\vec{a}\cdot c \approx \vec{b}$. In this formulation, the normal equations show $\vec{a}^\top\vec{a}\cdot c = \vec{a}^\top\vec{b}$, or

$$c = \frac{\vec{a}\cdot\vec{b}}{\vec{a}\cdot\vec{a}} = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|_2^2}.$$

We denote this *projection* of $\vec{b}$ onto $\vec{a}$ as:

$$\text{proj}_{\vec{a}}\,\vec{b} \equiv c\vec{a} = \frac{\vec{a}\cdot\vec{b}}{\vec{a}\cdot\vec{a}}\vec{a} = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|_2^2}\vec{a}$$

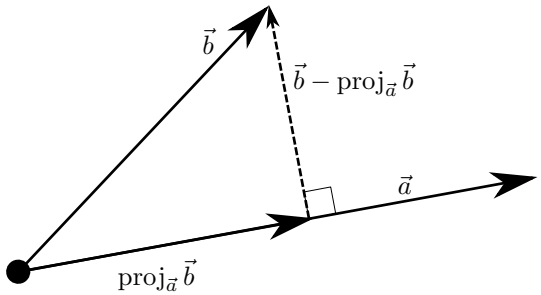


FIGURE 4.3 The projection $\text{proj}_{\vec{a}}\,\vec{b}$ is parallel to $\vec{a}$, while the remainder $\vec{b} - \text{proj}_{\vec{a}}\,\vec{b}$ is perpendicular to $\vec{a}$.

By design, $\text{proj}_{\vec{a}}\,\vec{b}$ is parallel to $\vec{a}$. What about the remainder $\vec{b} - \text{proj}_{\vec{a}}\vec{b}$? We can do the following computation to find out:

$$\begin{aligned}
\vec{a} \cdot (\vec{b} - \text{proj}_{\vec{a}}\,\vec{b}) &= \vec{a} \cdot \vec{b} - \vec{a} \cdot \left(\frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2^2} \vec{a} \right) \text{ by definition of } \text{proj}_{\vec{a}}\,\vec{b} \\
&= \vec{a} \cdot \vec{b} - \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|_2^2} (\vec{a} \cdot \vec{a}) \text{ by moving the constant outside the dot product} \\
&= \vec{a} \cdot \vec{b} - \vec{a} \cdot \vec{b} \text{ since } \vec{a} \cdot \vec{a} = \|\vec{a}\|_2^2 \\
&= 0
\end{aligned}$$

This simplification shows we have decomposed $\vec{b}$ into a component parallel to $\vec{a}$ and another component orthogonal to $\vec{a}$, as illustrated in Figure 4.3.

Now, suppose that $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_k$ are orthonormal; for clarity, in this section we will put hats over vectors with unit length. Then, for any single i by the projection formula above we know:

$$\text{proj}_{\hat{a}_i}\,\vec{b} = (\hat{a}_i \cdot \vec{b})\hat{a}_i$$

The denominator does not appear because $\|\hat{a}_i\|_2 = 1$ by definition. More generally, however, we can project $\vec{b}$ onto span $\{\hat{a}_1, \cdots, \hat{a}_k\}$ by minimizing the following energy function E over $c_1, \ldots, c_k \in \mathbb{R}$:

$$\begin{aligned}
E(c_1, c_2, \ldots, c_k) &\equiv \|c_1\hat{a}_1 + c_2\hat{a}_2 + \cdots + c_k\hat{a}_k - \vec{b}\|^2 \\
&= \left(\sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j (\hat{a}_i \cdot \hat{a}_j) \right) - 2\vec{b} \cdot \left(\sum_{i=1}^{k} c_i \hat{a}_i \right) + \vec{b} \cdot \vec{b} \\
&\qquad \text{by applying and expanding } \|\vec{v}\|_2^2 = \vec{v} \cdot \vec{v} \\
&= \sum_{i=1}^{k} \left(c_i^2 - 2c_i \vec{b} \cdot \hat{a}_i \right) + \|\vec{b}\|_2^2 \text{ since the } \hat{a}_i\text{'s are orthonormal}
\end{aligned}$$

The second step here is *only* valid because of orthonormality of the $\hat{a}_i$'s. At a minimum, the derivative of this energy with respect to c_i is zero for every i, yielding the relationship

$$0 = \frac{\partial E}{\partial c_i} = 2c_i - 2\vec{b} \cdot \hat{a}_i \implies c_i = \hat{a}_i \cdot \vec{b}.$$

```
function GRAM-SCHMIDT(v_1, v_2, ..., v_k)
  ▷ Computes an orthonormal basis â_1, ..., â_k for span{v_1, ..., v_k}
  ▷ Assumes v_1, ..., v_k are linearly independent.
  â_1 ← v_1/‖v_1‖_2                          ▷ Nothing to project out of the first vector
  for i ← 2, 3, ..., k
    p ← 0                                    ▷ Projection of v_i onto span{â_1, ..., â_{i-1}}
    for j ← 1, 2, ..., i − 1
      p ← p + (v_i · â_j)â_j                 ▷ Projecting onto orthonormal basis
    r ← v_i − p                              ▷ Residual is orthogonal to current basis
    â_i ← r/‖r‖_2                            ▷ Normalize this residual and add it to the basis
  return {â_1, ..., â_k}
```

FIGURE 4.4 The Gram-Schmidt algorithm for orthogonalization. This implementation assumes that the input vectors are linearly independent; in practice linearly dependence can be detected by checking for division by zero.

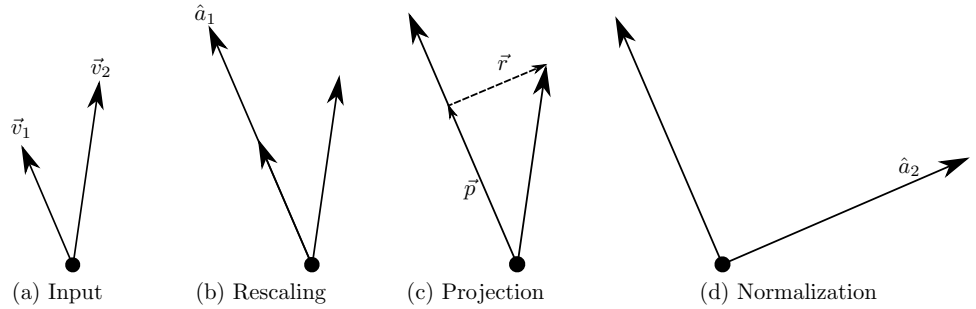


FIGURE 4.5 Steps of the Gram-Schmidt algorithm on two vectors $\vec{v}_1$ and $\vec{v}_2$ (a): $\hat{a}_1$ is a rescaled version of $\vec{v}_1$ (b); $\vec{v}_2$ is decomposed into a parallel component $\vec{p}$ and a residual $\vec{r}$ (c); $\vec{r}$ is normalized to obtain $\hat{a}_2$ (d).

Thus, we have shown that when $\hat{a}_1, \cdots, \hat{a}_k$ are orthonormal, the following relationship holds:

$$\text{proj}_{\text{span}\,\{\hat{a}_1,\cdots,\hat{a}_k\}}\,\vec{b} = (\hat{a}_1 \cdot \vec{b})\hat{a}_1 + \cdots + (\hat{a}_k \cdot \vec{b})\hat{a}_k$$

This formula extends the formula for $\text{proj}_{\vec{a}}\,\vec{b}$, and by a proof identical to the one above for single-vector projections, we must have

$$\hat{a}_i \cdot (\vec{b} - \text{proj}_{\text{span}\,\{\hat{a}_1,\cdots,\hat{a}_k\}}\,\vec{b}) = 0.$$

That is, we separated $\vec{b}$ into a component parallel to the span of the $\hat{a}_i$'s and a perpendicular residual.

4.4.2 Gram-Schmidt Algorithm

Our observations above lead to an algorithm for *orthogonalization*, or building an orthogonal basis $\{\hat{a}_1, \cdots, \hat{a}_k\}$ whose span is the same as that of a set of linearly independent but not necessarily orthogonal input vectors $\{\vec{v}_1, \cdots, \vec{v}_k\}$.

We add one vector at a time to the basis, starting with $\vec{v}_1$, then $\vec{v}_2$, and so on. When

```
function MODIFIED-GRAM-SCHMIDT(v_1, v_2, ..., v_k)
  ▷ Computes an orthonormal basis â_1, ..., â_k for span{v_1, ..., v_k}
  ▷ Assumes v_1, ..., v_k are linearly independent.

  for i ← 1, 2, ..., k
    â_i ← v_i / ||v_i||_2          ▷ Normalize the current vector and store in the basis
    for j ← i+1, i+2, ..., k
      v_j ← v_j − (v_j · â_i)â_i   ▷ Project â_i out of the remaining vectors
  return {â_1, ..., â_k}
```

FIGURE 4.6 The modified Gram-Schmidt algorithm.

add $\vec{v}_i$ is added to the current basis $\{\hat{a}_1, \ldots, \hat{a}_{i-1}\}$, we project out the span of $\hat{a}_1, \ldots, \hat{a}_{i-1}$. By the discussion in §4.4.1 the remaining residual must be orthogonal to the current basis, so we divide this residual by its norm to make it unit-length and add it to the basis. This technique, known as *Gram-Schmidt orthogonalization* is detailed in Figure 4.4 and illustrated in Figure 4.5.

If we start with a matrix $A \in \mathbb{R}^{m \times n}$ whose columns are $\vec{v}_1, \cdots, \vec{v}_k$, then we can implement Gram-Schmidt using a series of column operations on A. Dividing column i of A by its norm is equivalent to post-multiplying A by a $k \times k$ diagonal matrix. The projection step for column i involves subtracting only multiples of columns j with $j < i$, and thus this operation can be implemented with an upper-triangular elimination matrix. Thus, our discussion in §4.3 applies, and we can use Gram-Schmidt to obtain a factorization $A = QR$. When the columns of A are linearly independent, one way to find R is as a product $R = Q^\top A$; a more stable approach is to keep track of operations as we did for Gaussian elimination.

The Gram-Schmidt algorithm is well-known to be numerically unstable. There are many reasons for this instability that may or may not appear depending on the particular application. For instance, thanks to rounding and other issues, it might be the case that the $\hat{a}_i$'s are not completely orthogonal after the projection step. Our projection formula for finding $\vec{p}$ within the algorithm in Figure 4.4, however, only works when the $\hat{a}_i$'s are orthogonal. For this reason, in the presence of rounding the projection $\vec{p}$ of $\vec{v}_i$ becomes less accurate.

One way around this issue is the "modified Gram-Schmidt" (MGS) algorithm in Figure 4.6, which has similar running time but makes a subtle change in the way projections are computed. Rather than computing the projection $\vec{p}$ in each iteration i onto $\text{span}\{\hat{a}_1, \ldots, \hat{a}_{i-1}\}$, as soon as $\hat{a}_i$ is computed it is projected out of $\vec{v}_{i+1}, \ldots, \vec{v}_k$; subsequently we never have to consider $\hat{a}_i$ again. This way even if the basis globally is not completely orthogonal due to rounding, the projection step is valid since it only projects onto one $\hat{a}_i$ at a time. In the absence of rounding, modified Gram-Schmidt and classical Gram-Schmidt generate identical output.

A more subtle instability within the Gram-Schmidt algorithm that is not resolved by MGS can introduce serious numerical instabilities in the subtraction step. Suppose we provide the vectors $\vec{v}_1 = (1, 1)$ and $\vec{v}_2 = (1+\varepsilon, 1)$ as input to Gram-Schmidt for some $0 < \varepsilon \ll 1$. An reasonable basis for span $\{\vec{v}_1, \vec{v}_2\}$ is $\{(1, 0), (0, 1)\}$. But, if we apply Gram-Schmidt, we obtain:

$$\hat{a}_1 = \frac{\vec{v}_1}{\|\vec{v}_1\|} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$\vec{p} = \frac{2+\varepsilon}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

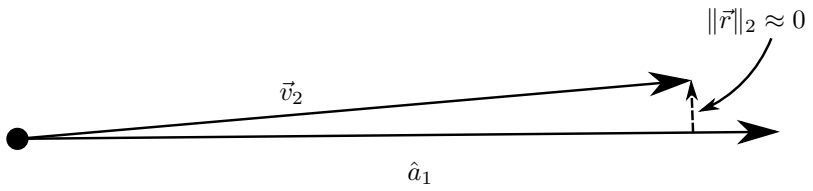


FIGURE 4.7 A failure mode of the basic and modified Gram-Schmidt algorithms; here $\hat{a}_1$ is nearly parallel to $\vec{v}_2$ and hence the residual $\vec{r}$ is vanishingly small.

$$\begin{aligned}\vec{r} = \vec{v}_2 - \vec{p} &= \begin{pmatrix} 1+\varepsilon \\ 1 \end{pmatrix} - \frac{2+\varepsilon}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} \varepsilon \\ -\varepsilon \end{pmatrix}\end{aligned}$$

Taking the norm, $\|\vec{v}_2 - \vec{p}_2\|_2 = (\sqrt{2}/2) \cdot \varepsilon$, so computing $\hat{a}_2$ will require division by a scalar on the order of ε. Division by small numbers is an unstable numerical operation that we should avoid. A similar case is illustrated in Figure 4.7.

4.5 HOUSEHOLDER TRANSFORMATIONS

In §4.3, we motivated the construction of QR factorization through the use of column operations. This construction is reasonable in the context of analyzing column spaces, but as we saw in our derivation of the Gram-Schmidt algorithm, the resulting numerical techniques can be unstable.

Rather than starting with A and post-multiplying by column operations to obtain $Q = AE_1 \cdots E_k$, however, we can preserve our high-level strategy from Gaussian elimination. That is, we can start with A and *pre*-multiply by orthogonal matrices Q_i to obtain $Q_k \cdots Q_1 A = R$; these Q's will act like row operations, eliminating elements of A until the resulting product R is upper-triangular. Then, thanks to orthogonality of the Q's we can write $A = (Q_1^\top \cdots Q_k^\top)R$, obtaining the QR factorization since products and transposes of orthogonal matrices are orthogonal.

The row operation matrices we used in Gaussian elimination and LU will not suffice for QR factorization since they are not orthogonal. Several alternatives have been suggested; we will introduce a common orthogonal row operation introduced in 1958 by Alston Scott Householder [60].

The space of orthogonal $n \times n$ matrices is very large, so we seek a smaller set of possible Q_i's that is easier to work with while still powerful enough to implement elimination operations. To develop some intuition, from our geometric discussions in §4.2 we know that orthogonal matrices must preserve angles and lengths, so intuitively they only can rotate and reflect vectors. Householder proposed using only reflection operations to reduce A to upper-triangular form; while we do not consider it here, a well-known alternative by Givens uses only rotations to accomplish the same task [44].

One way to write an orthogonal reflection matrix is in terms of projections, as illustrated in Figure 4.8. Suppose we have a vector $\vec{b}$ that we wish to reflect over a vector $\vec{v}$. We have shown that the residual $\vec{r} \equiv \vec{b} - \text{proj}_{\vec{v}}\,\vec{b}$ is perpendicular to $\vec{v}$. Following the reverse of this direction twice shows that the difference $2\text{proj}_{\vec{v}}\,\vec{b} - \vec{b}$ reflects $\vec{b}$ *over* $\vec{v}$.

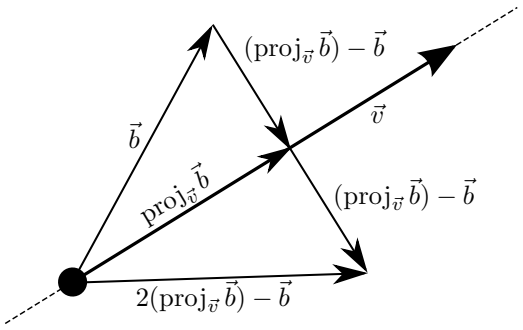


FIGURE 4.8 Reflecting $\vec{b}$ over $\vec{v}$.

We can expand our reflection formula as follows:

$$\begin{aligned}
2\text{proj}_{\vec{v}}\,\vec{b} - \vec{b} &= 2\frac{\vec{v}\cdot\vec{b}}{\vec{v}\cdot\vec{v}}\vec{v} - \vec{b} \text{ by definition of projection} \\
&= 2\vec{v}\cdot\frac{\vec{v}^\top\vec{b}}{\vec{v}^\top\vec{v}} - \vec{b} \text{ using matrix notation} \\
&= \left(\frac{2\vec{v}\vec{v}^\top}{\vec{v}^\top\vec{v}} - I_{n\times n}\right)\vec{b} \\
&\equiv -H_{\vec{v}}\vec{b}, \text{ where we define } H_v \equiv I_{n\times n} - \frac{2\vec{v}\vec{v}^\top}{\vec{v}^\top\vec{v}}
\end{aligned}$$

By this factorization, we can think of reflecting $\vec{b}$ over $\vec{v}$ as applying a linear operator $-H_{\vec{v}}$ to $\vec{b}$; $-H_{\vec{v}}$ has no dependence on $\vec{b}$. $H_{\vec{v}}$ without the negative is still orthogonal, and by convention we will use it from now on. Our derivation will parallel that in [54].

Like in forward substitution, in our first step we wish to pre-multiply A by a matrix that takes the first column of A, which we will denote $\vec{a}$, to some multiple of the first identity vector $\vec{e}_1$. Using reflections rather than forward substitutions, however, we now need to find some $\vec{v}, c$ such that $H_{\vec{v}}\vec{a} = c\vec{e}_1$. Expanding this relationship,

$$\begin{aligned}
c\vec{e}_1 &= H_{\vec{v}}\vec{a}, \text{ as explained above} \\
&= \left(I_{n\times n} - \frac{2\vec{v}\vec{v}^\top}{\vec{v}^\top\vec{v}}\right)\vec{a}, \text{ by definition of } H_{\vec{v}} \\
&= \vec{a} - 2\vec{v}\frac{\vec{v}^\top\vec{a}}{\vec{v}^\top\vec{v}}
\end{aligned}$$

Moving terms around shows

$$\vec{v} = (\vec{a} - c\vec{e}_1)\cdot\frac{\vec{v}^\top\vec{v}}{2\vec{v}^\top\vec{a}}$$

In other words, if $H_{\vec{v}}$ accomplishes the desired reflection then $\vec{v}$ must be parallel to the difference $\vec{a} - c\vec{e}_1$. Scaling $\vec{v}$ does not affect the formula for $H_{\vec{v}}$, so for now assuming such an $H_{\vec{v}}$ exists we can attempt to *choose* $\vec{v} = \vec{a} - c\vec{e}_1$.

If this choice is valid, then substituting $\vec{v} = \vec{a} - c\vec{e}_1$ into the simplified expression shows

$$\vec{v} = \vec{v}\cdot\frac{\vec{v}^\top\vec{v}}{2\vec{v}^\top\vec{a}}$$

Thus, assuming $\vec{v} \neq \vec{0}$, the coefficient next to $\vec{v}$ on the right hand side must be 1, showing:

$$\begin{aligned} 1 &= \frac{\vec{v}^\top \vec{v}}{2\vec{v}^\top \vec{a}} \\ &= \frac{\|\vec{a}\|_2^2 - 2c\vec{e}_1 \cdot a + c^2}{2(\vec{a} \cdot \vec{a} - c\vec{e}_1 \cdot \vec{a})} \\ \text{Or, } 0 = \|\vec{a}\|_2^2 - c^2 &\implies c = \pm\|\vec{a}\|_2 \end{aligned}$$

After choosing $c = \pm\|\vec{a}\|_2$, our steps above are all reversible. We set out to choose $\vec{v}$ such that $H_{\vec{v}}\vec{a} = c\vec{e}_1$. By taking $\vec{v} = \vec{a} - c\vec{e}_1$ and choosing $c = \pm\|\vec{a}\|_2$, the steps above show:

$$H_{\vec{v}}A = \begin{pmatrix} c & \times & \times & \times \\ 0 & \times & \times & \times \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \times & \times & \times \end{pmatrix}$$

We have just accomplished a step similar to forward elimination using orthogonal matrices!

To fully reduce A to upper triangular form, we must repeat the steps above to eliminate all elements of A below the diagonal. During the k-th step of triangularization, we can take $\vec{a}$ to be the k-th column of $Q_{k-1}Q_{k-2}\cdots Q_1 A$, where the Q_i's are reflection matrices like the one derived above. We can split $\vec{a}$ into two components:

$$\vec{a} = \begin{pmatrix} \vec{a}_1 \\ \vec{a}_2 \end{pmatrix}$$

Here, $\vec{a}_1 \in \mathbb{R}^{k-1}$ and $\vec{a}_2 \in \mathbb{R}^{m-k+1}$. We wish to find $\vec{v}$ such that

$$H_{\vec{v}}\vec{a} = \begin{pmatrix} \vec{a}_1 \\ c \\ \vec{0} \end{pmatrix}$$

Following a parallel derivation to the one above for the case $k = 1$ shows that

$$\vec{v} = \begin{pmatrix} \vec{0} \\ \vec{a}_2 \end{pmatrix} - c\vec{e}_k$$

accomplishes exactly this transformation when $c = \pm\|\vec{a}_2\|$. We usually choose the sign of c to avoid cancellation by making it have sign opposite to that of the k-th value in $\vec{a}$.

The algorithm for Householder QR, illustrated in Figure 4.9, applies these formulas. For each column of A, we compute $\vec{v}$ annihilating the bottom elements of the column and apply $H_{\vec{v}}$ to A. The end result is an upper triangular matrix $R = H_{\vec{v}_n}\cdots H_{\vec{v}_1}A$. Q is given by the product $H_{\vec{v}_1}^\top \cdots H_{\vec{v}_n}^\top$. When $m < n$, it may be preferable to store Q implicitly as a list of vectors $\vec{v}$, which fits in the lower triangle that otherwise would be empty in R.

4.6 REDUCED QR FACTORIZATION

We conclude our discussion by returning to the least-squares problem $A\vec{x} \approx \vec{b}$ when $A \in \mathbb{R}^{m\times n}$ is not square. Both algorithms we have discussed in this chapter can factor non-square matrices A into products QR, but the sizes of Q and R, resp., are somewhat different:

- When applying Gram-Schmidt, we do column operations on A to obtain Q by orthogonalization. For this reason, the dimension of A is that of Q, yielding $Q \in \mathbb{R}^{m\times n}$ and $R \in \mathbb{R}^{n\times n}$.

function HOUSEHOLDER-QR(A)
▷ Factors $A \in \mathbb{R}^{m\times n}$ as $A = QR$.
▷ $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{m\times n}$ is upper triangular

$Q \leftarrow I_{m\times m}$
$R \leftarrow A$
for $k \leftarrow 1, 2, \ldots, m$
$\vec{a} \leftarrow R\vec{e}_k$ ▷ Isolate column k of R and store it in $\vec{a}$
$(\vec{a}_1, \vec{a}_2) \leftarrow$ SPLIT($\vec{a}$,$k-1$) ▷ Separate off the first $k-1$ elements of $\vec{a}$
$c \leftarrow \|\vec{a}_2\|_2$ ▷ Find reflection vector $\vec{v}$ for the Householder matrix $H_{\vec{v}}$
$\vec{v} \leftarrow \begin{pmatrix} \vec{0} \\ \vec{a}_2 \end{pmatrix} - c\vec{e}_k$
$R \leftarrow H_{\vec{v}}R$ ▷ Eliminate elements below the diagonal of the k-th column
$Q \leftarrow QH_{\vec{v}}^{\top}$
return Q, R

FIGURE 4.9 Householder QR factorization; the products with $H_{\vec{v}}$ can be carried out in quadratic time after expanding the particular formula for $H_{\vec{v}}$ in terms of $\vec{v}$.

- When using Householder reflections, we obtain Q as the product of a number of $m \times m$ reflection matrices, leaving $R \in \mathbb{R}^{m\times n}$.

Suppose we are in the typical case for least-squares, for which $m \gg n$. We still prefer to use the Householder method due to its numerical stability, but now the $m \times m$ matrix Q might be too large to store. To save space, we can use the upper triangular structure of R to our advantage. For instance, consider the structure of a 5×3 matrix R:

$$R = \left(\begin{array}{ccc} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right)$$

Anything below the upper $n \times n$ square of R must be zero, yielding a simplification:

$$A = QR = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1 R_1$$

Here, $Q_1 \in \mathbb{R}^{m\times n}$ and $R_1 \in \mathbb{R}^{n\times n}$ still contains the upper triangle of R. This is called the "reduced" QR factorization of A, since the columns of Q_1 contain a basis for the column space of A rather than for all of $\mathbb{R}^m$; it takes up far less space. The discussion in §4.3 still applies, so the reduced QR factorization can be used for least-squares in a similar fashion.

4.7 EXERCISES

4.1 Suppose $A \in \mathbb{R}^{m\times n}$ is factored $A = QR$. Show that $P_0 = I_{m\times m} - QQ^{\top}$ is the projection matrix onto the null space of $A^{\top}$.

4.2 Suppose we consider $\vec{a} \in \mathbb{R}^n$ as an $n \times 1$ matrix. Write out its "reduced" QR factorization explicitly.

4.3 Propose a method for finding the least-norm projection of a vector $\vec{v}$ onto the column space of $A \in \mathbb{R}^{m \times n}$ with $m > n$.

4.4 Alternatives to the QR factorization:

(a) Can a matrix $A \in \mathbb{R}^{m \times n}$ be factored into $A = RQ$ where R is upper triangular and Q is orthogonal? How?

(b) Can a matrix $A \in \mathbb{R}^{m \times n}$ be factored into $A = QL$ where L is lower triangular?

4.5 Relating QR and Cholesky factorizations:

(a) Take $A \in \mathbb{R}^{m \times n}$ and suppose we apply the Cholesky factorization to obtain $A^\top A = LL^\top$. Define $Q \equiv A(L^\top)^{-1}$. Show that Q is orthogonal.

(b) Based on the previous part, suggest a relationship between the Cholesky factorization of $A^\top A$ and QR factorization of A.

4.6 Suppose $A \in \mathbb{R}^{m \times n}$ is rank m with $m < n$. Suppose we factor

$$A^\top = Q \begin{pmatrix} R_1 \\ 0 \end{pmatrix}.$$

Provide a solution $\vec{x}$ to the underdetermined system $A\vec{x} = \vec{b}$ in terms of Q and R_1.

Hint: Try the square case $A \in \mathbb{R}^{n \times n}$ first, and use the result to guess a form for $\vec{x}$. Be careful that you multiply matrices of proper size.

4.7 ("Generalized QR," [2]) One way to generalize the QR factorization of a matrix is to consider the possibility of factorizing multiple matrices *simultaneously*.

(a) Suppose $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times p}$, with $m \leq n \leq p$. Show that there are orthogonal matrices $Q \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ as well as a matrix $R \in \mathbb{R}^{n \times m}$ such that the following conditions hold:

- $Q^\top A = R$
- $Q^\top BV = S$, where S can be written

$$S = \begin{pmatrix} 0 & \bar{S} \end{pmatrix},$$

for upper-triangular $\bar{S} \in \mathbb{R}^{n \times n}$

- R can be written

$$R = \begin{pmatrix} \bar{R} \\ 0 \end{pmatrix},$$

for upper-triangular $\bar{R} \in \mathbb{R}^{m \times m}$

Hint: Take $\bar{R}$ to be R_1 from the reduced QR factorization of A. Apply RQ factorization to $Q^\top B$; see problem 4.4a.

(b) Show how to solve the following optimization problem for $\vec{x}$ and $\vec{u}$ using the generalized QR factorization:

$$\begin{array}{ll} \min_{\vec{x},\vec{u}} & \|\vec{u}\|_2 \\ \text{such that} & A\vec{x} + B\vec{u} = \vec{c} \end{array}$$

You can assume $\bar{S}$ and $\bar{R}$ are invertible.

4.8 An alternative algorithm for QR factorization uses *Givens rotations* rather than Householder reflections.

(a) The 2×2 rotation matrix by angle $\theta \in [0, 2\pi)$ is given by

$$R_\theta \equiv \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

Show that for a given $\vec{x} \in \mathbb{R}^2$, a θ always exists such that $R_\theta \vec{x} = r\vec{e}_1$, where $r \in \mathbb{R}$ and $\vec{e}_1 = (1, 0)$. Give formulas for $\cos\theta$ and $\sin\theta$ that do not require trigonometric functions.

(b) The *Givens rotation matrix* of rows i and k about angle θ is given by

$$G(i, j, \theta) \equiv \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix},$$

where $c \equiv \cos\theta$ and $s \equiv \sin\theta$. In this formula, the c's appear in positions (i, i) and (j, j) while the s's appear in positions (i, j) and (j, i). Provide an $O(n)$ method for finding the product $G(i, j, \theta)A$ for $A \in \mathbb{R}^{n \times n}$; the matrix A can be modified in the process.

(c) Give an $O(n^3)$ time algorithm for overwriting $A \in \mathbb{R}^{n \times n}$ with $Q^\top A = R$, where $Q \in \mathbb{R}^{n \times n}$ is orthogonal and $R \in \mathbb{R}^{n \times n}$ is upper-triangular. You do not need to store Q.

(d) Suggest how you might store Q implicitly if you use the QR method you developed in the previous part.

(e) Suggest an $O(n^3)$ method for recovering the matrix Q given A and R.

4.9 (adapted from [46], §5.1) If $\vec{x}, \vec{y} \in \mathbb{R}^m$ with $\|\vec{x}\|_2 = \|\vec{y}\|_2$, write an algorithm for finding an orthogonal matrix Q such that $Q\vec{x} = \vec{y}$.

4.10 ("TSQR," [26]) The QR factorization algorithms we considered can be challenging to extend to parallel architectures like MapReduce. Here, we consider QR factorization of $A \in \mathbb{R}^{m \times n}$ where $m \gg n$.

(a) Suppose $A \in \mathbb{R}^{8n \times n}$. Factor $A = Q\bar{R}$, where $Q \in \mathbb{R}^{8n \times 8n}$ is orthogonal and $\bar{R} \in \mathbb{R}^{4n \times n}$ contains four $n \times n$ upper triangular blocks.

(b) Recursively apply your answer from 4.10a to generate a QR factorization of A.

(c) Now, write

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{pmatrix}.$$

Suppose we make the following factorizations:

$$\begin{aligned} A_1 &= Q_1 R_1 \\ \begin{pmatrix} R_1 \\ A_2 \end{pmatrix} &= Q_2 R_2 \\ \begin{pmatrix} R_2 \\ A_3 \end{pmatrix} &= Q_3 R_3 \\ \begin{pmatrix} R_3 \\ A_4 \end{pmatrix} &= Q_4 R_4, \end{aligned}$$

where each of the R_i's are square. Use these matrices to factor $A = QR$.

(d) Suppose we read A row-by-row. Why might the simplification in 4.10c be useful for QR factorization of A in this case?

CHAPTER 5

Eigenvectors

CONTENTS

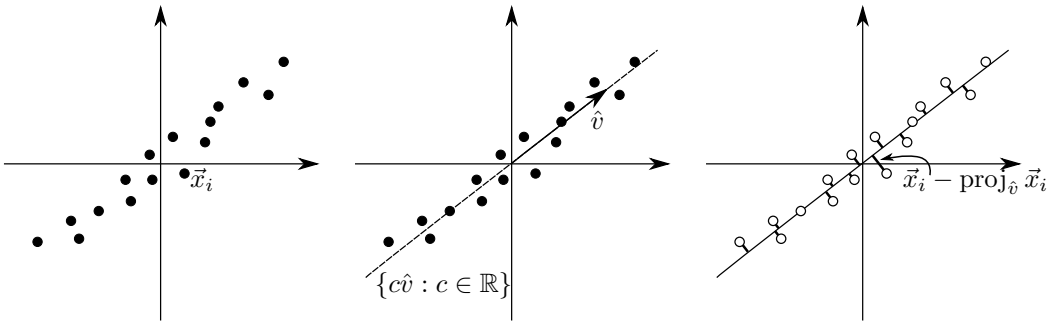
5.1 Motivation 101
5.1.1 Statistics 102
5.1.2 Differential Equations 103
5.1.3 Spectral Embedding 104
5.2 Properties of Eigenvectors 106
5.2.1 Symmetric and Positive Definite Matrices 108
5.2.2 Specialized Properties 110
5.2.2.1 Characteristic Polynomial 110
5.2.2.2 Jordan Normal Form 110
5.3 Computing A Single Eigenvalue 111
5.3.1 Power Iteration 111
5.3.2 Inverse Iteration 112
5.3.3 Shifting 113
5.4 Finding Multiple Eigenvalues 114
5.4.1 Deflation 114
5.4.2 QR Iteration 115
5.4.3 Krylov Subspace Methods 119
5.5 Sensitivity and Conditioning 120

WE turn our attention now to a *nonlinear* problem about matrices: Finding their eigenvalues and eigenvectors. Eigenvectors $\vec{x}$ and their corresponding eigenvalues λ of a square matrix A are determined by the equation $A\vec{x} = \lambda\vec{x}$. There are many ways to see that this problem is nonlinear. For instance, there is a *product* of unknowns λ and $\vec{x}$, and to avoid the trivial solution $\vec{x} = \vec{0}$ we constrain $\|\vec{x}\|_2 = 1$; this constraint is circular rather than linear. Thanks to this structure, our methods for finding eigenspaces will be considerably different from techniques for solving and analyzing linear systems of equations.

5.1 MOTIVATION

Despite the arbitrary-looking form of the equation $A\vec{x} = \lambda\vec{x}$, the problem of finding eigenvectors and eigenvalues of a matrix A arises naturally in many circumstances. To illustrate this point, before presenting algorithms for finding eigenvectors and eigenvalues we motivate our discussion with a few examples below.

It is worth reminding ourselves of one source of eigenvalue problems already considered in Chapter 0. As explained in Example 0.27, the following fact will guide many of our modeling decisions:



(a) Input data (b) Principal axis (c) Projection error

FIGURE 5.1 (a) A dataset with correlation between the horizontal and vertical axes; (b) we seek the unit vector $\hat{v}$ such that all data points are well-approximated by some point along $\text{span}\{\hat{v}\}$; (c) to find $\hat{v}$, we can minimize the sum of squared residual norms $\sum_i \|\vec{x}_i - \text{proj}_{\hat{v}}\, \vec{x}_i\|_2^2$ with the constraint that $\|\hat{v}\|_2 = 1$.

When A is symmetric, the eigenvectors of A are the critical points of $\vec{x}^\top A\vec{x}$ with the constraint $\|\vec{x}\|_2 = 1$.

A theme common to many of our example eigenvalue problems is this optimization or a similar one minimizing $\|A\vec{x}\|_2^2 = \vec{x}^\top(A^\top A)\vec{x}$, providing a clue for how to model problems using eigenspaces.

5.1.1 Statistics

Suppose we have machinery for collecting several statistical observations about a collection of items. For instance, in a medical study we may collect the age, weight, blood pressure, and heart rate of every patient in a hospital. Then, each patient i can be represented by a point $\vec{x}_i \in \mathbb{R}^4$ storing these four values.

These statistics may exhibit strong correlations between the different dimensions, as in Figure 5.1(a). For instance, patients with higher blood pressure may be likely to have higher weights or heart rates. For this reason, although we collected our data in $\mathbb{R}^4$, in reality it may—to some approximate degree—live in a lower-dimensional space better capturing the relationships between the different dimensions.

For now, suppose that there exists a *one*-dimensional space approximating our dataset, illustrated in Figure 5.1(b). Then, we expect all the data points to be nearly parallel to some vector $\vec{v}$, so that each point $\vec{x}_i$ can be written as $\vec{x}_i \approx c_i\vec{v}$ for different $c_i \in \mathbb{R}$. From before, we know that the best approximation of $\vec{x}_i$ parallel to $\vec{v}$ is $\text{proj}_{\vec{v}}\, \vec{x}_i$. Defining $\hat{v} \equiv \vec{v}/\|\vec{v}\|$, we can write

$$\begin{aligned}\text{proj}_{\vec{v}}\, \vec{x}_i &= \frac{\vec{x}_i \cdot \vec{v}}{\vec{v} \cdot \vec{v}}\vec{v} \text{ by definition} \\ &= (\vec{x}_i \cdot \hat{v})\hat{v} \text{ since } \vec{v} \cdot \vec{v} = \|\vec{v}\|_2^2.\end{aligned}$$

The magnitude of $\vec{v}$ does not matter for the problem at hand, since the projection of $\vec{x}_i$ onto any nonzero multiple of $\hat{v}$ is the same, so it is reasonable to restrict our search to the space of *unit* vectors $\hat{v}$.

Following the pattern of least squares, we have a new optimization problem:

$$\begin{aligned}&\text{minimize}_{\hat{v}} \sum_i \|\vec{x}_i - \text{proj}_{\hat{v}} \vec{x}_i\|_2^2\\ &\text{such that } \|\hat{v}\|_2 = 1\end{aligned}$$

This problem minimizes the sum of squared differences between the data points $\vec{x}_i$ and their best approximation as a multiple of $\hat{v}$, as in Figure 5.1(c). We can simplify our optimization objective using the observations we already have made and some linear algebra:

$$\begin{aligned}\sum_i \|\vec{x}_i - \text{proj}_{\hat{v}} \vec{x}_i\|_2^2 &= \sum_i \|\vec{x}_i - (\vec{x}_i \cdot \hat{v})\hat{v}\|_2^2 \text{ as explained above}\\ &= \sum_i \left(\|\vec{x}_i\|_2^2 - 2(\vec{x}_i \cdot \hat{v})(\vec{x}_i \cdot \hat{v}) + (\vec{x}_i \cdot \hat{v})^2\|\hat{v}\|_2^2\right) \text{ since } \|\vec{w}\|_2^2 = \vec{w} \cdot \vec{w}\\ &= \sum_i \left(\|\vec{x}_i\|_2^2 - (\vec{x}_i \cdot \hat{v})^2\right) \text{ since } \|\hat{v}\|_2 = 1\\ &= \text{const.} - \sum_i (\vec{x}_i \cdot \hat{v})^2 \text{ since the unknown here is } \hat{v}\\ &= \text{const.} - \|X^\top \hat{v}\|_2^2, \text{ where the columns of } X \text{ are the vectors } \vec{x}_i.\end{aligned}$$

After removing the negative sign, this derivation shows that we can solve an equivalent maximization problem:

$$\begin{aligned}&\text{maximize } \|X^\top \hat{v}\|_2^2\\ &\text{such that } \|\hat{v}\|_2^2 = 1,\end{aligned}$$

Statisticians may recognize this equivalence as maximizing variance rather than minimizing approximation error.

We know $\|X^\top \hat{v}\|_2^2 = \hat{v}^\top X X^\top \hat{v}$, so by Example 0.27, $\hat{v}$ is exactly the eigenvector of $XX^\top$ with the highest eigenvalue. The vector $\hat{v}$ is known as the first *principal component* of the dataset.

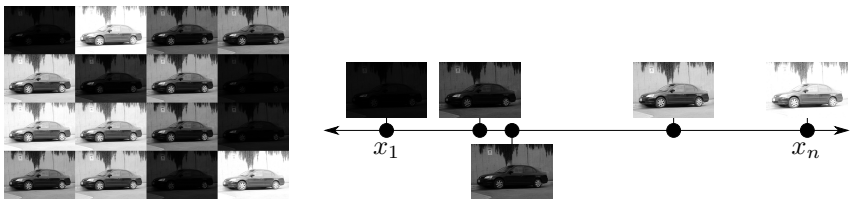
5.1.2 Differential Equations

Many physical forces can be written as functions of position. For instance, the force between two particles at positions $\vec{x}$ and $\vec{y}$ in $\mathbb{R}^3$ exerted by a spring can be written as $k(\vec{x} - \vec{y})$ by Hooke's Law; such spring forces are used to approximate forces holding cloth together in many simulation systems. Although forces are not always *linear* in position, we often approximate them in a linear fashion. In particular, in a physical system with n particles, we can encode the positions of all the particles simultaneously in a vector $\vec{X} \in \mathbb{R}^{3n}$. Then, the forces in the system can be approximated as $\vec{F} \approx A\vec{X}$ for some matrix A.

Newton's second law of motion states $F = ma$, or force equals mass times acceleration. In our context, we can write a diagonal *mass matrix* $M \in \mathbb{R}^{3n \times 3n}$ containing the mass of each particle in the system. Then, we know $\vec{F} = M\vec{X}''$, where prime denotes differentiation in time. By definition, $\vec{X}'' = (\vec{X}')'$, so after defining $\vec{V} \equiv \vec{X}'$ we have a *first*-order system of equations:

$$\frac{d}{dt}\begin{pmatrix}\vec{X}\\ \vec{V}\end{pmatrix} = \begin{pmatrix}0 & I_{3n\times 3n}\\ M^{-1}A & 0\end{pmatrix}\begin{pmatrix}\vec{X}\\ \vec{V}\end{pmatrix}$$

Here, we simultaneously compute both positions in $\vec{X} \in \mathbb{R}^{3n}$ and velocities $\vec{V} \in \mathbb{R}^{3n}$ of all n particles as functions of time; we will explore this reduction in more detail in Chapter 14.



(a) Database of photos (b) Spectral embedding

FIGURE 5.2 Suppose we are given an unsorted database of photographs (a) with some matrix W measuring the similarity between image i and image j. The one-dimensional spectral embedding (b) assigns each photograph i a value x_i so that if images i and j are similar then x_i will be close to x_j. Figure generated by D. Hyde

Beyond this reduction, differential equations of the form $\vec{x}' = A\vec{x}$ for an unknown function $\vec{x}(t)$ and fixed matrix A appear in many contexts, including simulation of cloth, springs, heat, waves, and other phenomena. Suppose we know eigenvectors $\vec{x}_1, \ldots, \vec{x}_k$ of A, such that $A\vec{x}_i = \lambda_i \vec{x}_i$. If we write the initial condition of the differential equation in terms of the eigenvectors, as

$$\vec{x}(0) = c_1\vec{x}_1 + \cdots + c_k\vec{x}_k,$$

then the solution to the equation can be written in closed form:

$$\vec{x}(t) = c_1 e^{\lambda_1 t}\vec{x}_1 + \cdots + c_k e^{\lambda_k t}\vec{x}_k.$$

That is, if we write the initial conditions of this differential equation in terms of the eigenvectors of A, then we know its solution for all times $t \geq 0$ for free; in problem 5.1 you will check this formula. This trick is not the end of the story for simulation: Finding the complete set of eigenvectors of A is expensive, and A may change over time.

5.1.3 Spectral Embedding

Suppose we have a collection of n items in a dataset and a measure $w_{ij} \geq 0$ of how similar elements i and j are; by symmetry, we will assume $w_{ij} = w_{ji}$. For instance, maybe we are given a collection of photographs as in Figure 5.2(a) and take w_{ij} to be a measure of the amount of overlap between the distributions of colors in photo i and in photo j.

Given the matrix W of w_{ij} values, we might wish to sort the photographs based on their similarity to simplify viewing and exploring the collection. That is, we could lay them out on a line so that the pair of photos i and j is close when w_{ij} is large, as in Figure 5.2(b). The measurements in w_{ij} may be noisy or inconsistent, however, so it may not be obvious how to sort the n photos using the n^2 values in W directly.

One way to order the collection would be to assign a number x_i to each item i, asking that similar objects are assigned similar numbers; we will then sort the collection based on the values in $\vec{x}$. We can measure how well an assignment of values in $\vec{x}$ groups similar objects by using the energy function

$$E(\vec{x}) \equiv \sum_{ij} w_{ij}(x_i - x_j)^2.$$

The difference $(x_i - x_j)^2$ is small when x_i and x_j receive similar values. Hence, given the

weighting w_{ij} next to $(x_i - x_j)^2$, minimizing $E(\vec{x})$ asks that items i and j with high similarity scores w_{ij} get mapped the closest.

Minimizing $E(\vec{x})$ with no constraints gives a minimum $\vec{x}$ with $E(\vec{x}) = 0$: $x_i = \text{const.}$ for all i. Furthermore, adding a constraint $\|\vec{x}\|_2 = 1$ does *not* remove this constant solution: Taking $x_i = 1/\sqrt{n}$ for all i gives $\|\vec{x}\|_2 = 1$ and $E(\vec{x}) = 0$. Thus, we must remove this case as well:

$$\begin{aligned} &\text{minimize } E(\vec{x}) \\ &\text{such that } \|\vec{x}\|_2^2 = 1 \\ &\qquad \vec{1} \cdot \vec{x} = 0 \end{aligned}$$

Our second constraint requires that the sum of $\vec{x}$ is zero, preventing $x_1 = x_2 = \cdots = x_n$ when combined with the $\|\vec{x}\|_2 = 1$ constraint.

We can simplify the energy in a few steps:

$$\begin{aligned} E(\vec{x}) &= \sum_{ij} w_{ij}(x_i - x_j)^2 \\ &= \sum_{ij} w_{ij}(x_i^2 - 2x_i x_j + x_j^2) \\ &= \sum_i a_i x_i^2 - 2\sum_{ij} w_{ij} x_i x_j + \sum_j a_j x_j^2 \text{ where } \vec{a} \equiv W\vec{1}, \text{ since } W^\top = W \\ &= 2\vec{x}^\top (A - W)\vec{x} \text{ where } A = \text{diag}(\vec{a}) \end{aligned}$$

We can check that $\vec{1}$ is an eigenvector of $A - W$ with eigenvalue 0:

$$(A - W)\vec{1} = A\vec{1} - W\vec{1} = \vec{a} - \vec{a} = \vec{0}.$$

More interestingly, the eigenvector corresponding to the *second*-smallest eigenvalue is the minimizer for our constrained problem above! One way to see this fact is to write the Lagrange multiplier function corresponding to this optimization:

$$\Lambda \equiv 2\vec{x}^\top (A - W)\vec{x} - \lambda(1 - \|\vec{x}\|_2^2) - \mu(\vec{1} \cdot \vec{x})$$

Applying Theorem 0.1, at the optimal point we must have:

$$\begin{aligned} 0 &= \nabla_{\vec{x}} \Lambda = 4(A - W)\vec{x} - 2\lambda\vec{x} - \mu\vec{1} \\ 1 &= \|\vec{x}\|_2^2 \\ 0 &= \vec{1} \cdot \vec{x} \end{aligned}$$

If we take the dot product of both sides of the first expression with $\vec{1}$, we find:

$$\begin{aligned} 0 &= \vec{1} \cdot [4(A - W)\vec{x} - 2\lambda\vec{x} - \mu\vec{1}] \\ &= 4\vec{1}^\top (A - W)\vec{x} - \mu n \text{ where } \vec{x} \in \mathbb{R}^n, \text{ since } \vec{1} \cdot \vec{x} = 0 \\ &= -\mu n \text{ since } A\vec{1} = W\vec{1} = \vec{a} \\ \implies \mu &= 0. \end{aligned}$$

Substituting this new observation into the Lagrange multiplier condition, we find:

$$2(A - W)\vec{x} = \lambda\vec{x}$$

We explicitly ignore the eigenvalue $\lambda = 0$ of $A - W$ corresponding to the eigenvector $\vec{1}$, so $\vec{x}$ must be the eigenvector with *second*-smallest eigenvalue. The resulting $\vec{x}$ is the "spectral embedding" of W onto one dimension, referring to the fact that we call the set of eigenvalues of a matrix its spectrum. Taking more eigenvectors of $A-W$ provides embeddings into higher dimensions.

5.2 PROPERTIES OF EIGENVECTORS

We have established a variety of applications in need of eigenspace computation. Before we can explore algorithms for this purpose, however, we will more closely examine the structure of the eigenvalue problem.

We can begin with a few definitions that likely are evident at this point:

Definition 5.1 (Eigenvalue and eigenvector)**.** An *eigenvector* $\vec{x} \neq \vec{0}$ of a matrix $A \in \mathbb{R}^{n\times n}$ is any vector satisfying $A\vec{x} = \lambda\vec{x}$ for some $\lambda \in \mathbb{R}$; the corresponding λ is known as an *eigenvalue*. *Complex* eigenvalues and eigenvectors satisfy the same relationships with $\lambda \in \mathbb{C}$ and $\vec{x} \in \mathbb{C}^n$.

Definition 5.2 (Spectrum and spectral radius)**.** The *spectrum* of A is the set of eigenvalues of A. The *spectral radius* $\rho(A)$ is the eigenvalue λ maximizing $|\lambda|$.

The scale of an eigenvector is not important. In particular, we can check $A(c\vec{x}) = cA\vec{x} = c\lambda\vec{x} = \lambda(c\vec{x})$, so $c\vec{x}$ is an eigenvector with the same eigenvalue. For this reason, we can restrict our search to those eigenvectors $\vec{x}$ with $\|\vec{x}\|_2 = 1$ without losing any nontrivial structure. Adding this constraint does not completely relieve ambiguity, since $\pm\vec{x}$ are both eigenvectors with the same eigenvalue, but this case is easier to detect and deal with.

The algebraic properties of eigenvectors and eigenvalues are the subject of many mathematical studies in themselves. Thankfully, a few basic properties will suffice for the discussion at hand, and hence we will limit our discussion to a few theorems that affect the design of numerical algorithms; we parallel the development of [4].

First, we should check that every matrix has at least one eigenvector so that our search is not in vain. Our strategy for this and other related problems is to notice that λ is an eigenvalue such that $A\vec{x} = \lambda\vec{x}$ if and only if $(A - \lambda I_{n\times n})\vec{x} = \vec{0}$; in other words, λ is an eigenvalue exactly when the matrix $A - \lambda I_{n\times n}$ is not full-rank.

Proposition 5.1 ([4], Theorem 2.1)**.** Every matrix $A \in \mathbb{R}^{n\times n}$ has at least one (potentially complex) eigenvector.

Proof. Take any vector $\vec{x} \in \mathbb{R}^n \backslash \{\vec{0}\}$. The set $\{\vec{x}, A\vec{x}, A^2\vec{x}, \cdots, A^n\vec{x}\}$ must be linearly dependent because it contains $n+1$ vectors in n dimensions. So, there exist constants $c_0, \ldots, c_n \in \mathbb{R}$ not all zero such that $\vec{0} = c_0\vec{x} + c_1A\vec{x} + \cdots + c_nA^n\vec{x}$. We define a polynomial

$$f(z) \equiv c_0 + c_1z + \cdots + c_nz^n.$$

By the Fundamental Theorem of Algebra, there exist $m \geq 1$ roots $z_i \in \mathbb{C}$ and $c \neq 0$ such that

$$f(z) = c_n(z - z_1)(z - z_2)\cdots(z - z_m).$$

Applying this factorization, we can write:

$$\begin{aligned}\vec{0} &= c_0\vec{x} + c_1A\vec{x} + \cdots + c_nA^n\vec{x} \\ &= (c_0I_{n\times n} + c_1A + \cdots + c_nA^n)\vec{x} \\ &= c_n(A - z_1I_{n\times n})\cdots(A - z_mI_{n\times n})\vec{x}.\end{aligned}$$

In this form, it is clear that at least one $A - z_i I_{n\times n}$ has a null space, showing that there exists $\vec{v}$ with $A\vec{v} = z_i\vec{v}$, as needed. □

There is one additional fact worth checking to motivate our discussion of eigenvector computation. While it can be the case that a single eigenvalue admits more than one corresponding eigenvector, when two eigenvectors have different eigenvalues they cannot be related in the following sense:

Proposition 5.2 ([4], Proposition 2.2)**.** Eigenvectors corresponding to different eigenvalues must be linearly independent.

Proof. Suppose this is not the case. Then there exist eigenvectors $\vec{x}_1, \ldots, \vec{x}_k$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ that are linearly dependent. This implies that there are coefficients $c_1, \ldots, c_k$ not all zero with $\vec{0} = c_1\vec{x}_1 + \cdots + c_k\vec{x}_k$.

For any two indices i and j, since $A\vec{x}_j = \lambda_j\vec{x}_j$, we can simplify the product

$$(A - \lambda_i I_{n\times n})\vec{x}_j = A\vec{x}_j - \lambda_i\vec{x}_j = \lambda_j\vec{x}_j - \lambda_i\vec{x}_j = (\lambda_j - \lambda_i)\vec{x}_j.$$

Hence, if we premultiply the relationship $\vec{0} = c_1\vec{x}_1 + \cdots + c_k\vec{x}_k$ by the matrix $(A - \lambda_2 I_{n\times n})\cdots(A - \lambda_k I_{n\times n})$, we find:

$$\begin{aligned}\vec{0} &= (A - \lambda_2 I_{n\times n})\cdots(A - \lambda_k I_{n\times n})(c_1\vec{x}_1 + \cdots + c_k\vec{x}_k)\\ &= c_1(\lambda_1 - \lambda_2)\cdots(\lambda_1 - \lambda_k)\vec{x}_1.\end{aligned}$$

Since all the λ_i's are distinct, this shows $c_1 = 0$. The same argument shows that the rest of the c_i's have to be zero, contradicting linear dependence. □

This proposition shows that an $n\times n$ matrix can have at most n distinct eigenvalues, since a set of n eigenvalues yields n linearly independent vectors. The maximum number of linearly independent eigenvectors corresponding to an eigenvalue λ is the *geometric multiplicity* of λ. It is not true, however, that a matrix has to have *exactly* n linearly independent eigenvectors. This is the case for many matrices, which we will call *nondefective*:

Definition 5.3 (Nondefective)**.** A matrix $A \in \mathbb{R}^{n\times n}$ is *nondefective* or *diagonalizable* if its eigenvectors span $\mathbb{R}^n$.

Example 5.1 (Defective matrix)**.** The matrix

$$\begin{pmatrix} 5 & 2 \\ 0 & 5 \end{pmatrix}$$

has only one eigenvector $(1, 0)$.

We call nondefective matrices *diagonalizable* for the following reason: If a matrix is diagonalizable, then it has n eigenvectors $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^n$ with corresponding (possibly non-unique) eigenvalues $\lambda_1, \ldots, \lambda_n$. Take the columns of X to be the vectors $\vec{x}_i$, and define D to be the diagonal matrix with $\lambda_1, \ldots, \lambda_n$ along the diagonal. Then, we have $AX = XD$; this relationship is a "stacked" version of $A\vec{x}_i = \lambda_i\vec{x}_i$. In other words, $D = X^{-1}AX$, meaning A is diagonalized by a *similarity transformation* $A \mapsto X^{-1}AX$:

Definition 5.4 (Similar matrices)**.** Two matrices A and B are *similar* if there exists T with $B = T^{-1}AT$.

Similar matrices have the same eigenvalues, since if $B\vec{x} = \lambda x$, then $T^{-1}AT\vec{x} = \lambda\vec{x}$; hence, $A(T\vec{x}) = \lambda(T\vec{x})$, showing $T\vec{x}$ is an eigenvector of A with eigenvalue λ. In other words:

We can apply all the similarity transformations we want to a matrix without modifying its set of eigenvalues.

This observation is the foundation of many eigenvector computation methods, which start with a general matrix A and reduce it to a matrix whose eigenvalues are more obvious by applying similarity transformations; this procedure is analogous to applying row operations to reduce a matrix to triangular form.

5.2.1 Symmetric and Positive Definite Matrices

Unsurprisingly given our special consideration of normal matrices $A^\top A$ in previous chapters, symmetric and/or positive definite matrices enjoy special eigenvector structure. If we can verify *a priori* that a matrix is symmetric or positive definite, specialized algorithms can be used to extract its eigenvectors more quickly.

Our original definition of eigenvalues allows them to be complex values in $\mathbb{C}$ even if A is a real matrix. We can prove, however, that in the symmetric case we do not need complex arithmetic. To do so, we will generalize symmetric matrices to matrices in $\mathbb{C}^{n\times n}$ by introducing the set of *Hermitian* matrices:

Definition 5.5 (Complex conjugate)**.** The *complex conjugate* of a number $z = a+bi \in \mathbb{C}$, where $a, b \in \mathbb{R}$, is $\bar{z} \equiv a - bi$.

Definition 5.6 (Conjugate transpose)**.** The *conjugate transpose* of $A \in \mathbb{C}^{m\times n}$ is $A^H \equiv \bar{A}^\top$.

Definition 5.7 (Hermitian matrix)**.** A matrix $A \in \mathbb{C}^{n\times n}$ is *Hermitian* if $A = A^H$.

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is automatically Hermitian because it has no complex part.

We also can generalize the notion of a dot product to complex vectors by defining an *inner product* as follows:

$$\langle \vec{x}, \vec{y} \rangle \equiv \sum_i x_i \bar{y}_i,$$

where $\vec{x}, \vec{y} \in \mathbb{C}^n$. Once again this definition coincides with $\vec{x} \cdot \vec{y}$ when $\vec{x}, \vec{y} \in \mathbb{R}^n$; in the complex case, however, dot product symmetry is replaced by the condition $\langle \vec{v}, \vec{w} \rangle = \overline{\langle \vec{w}, \vec{v} \rangle}$.

We now can prove that it is not necessary to search for complex eigenvalues of symmetric or Hermitian matrices:

Proposition 5.3. All eigenvalues of Hermitian matrices are real.

Proof. Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian with $A\vec{x} = \lambda\vec{x}$. By scaling we can assume $\|\vec{x}\|^2 = \langle \vec{x}, \vec{x} \rangle = 1$. Then:

$$\begin{aligned}
\lambda &= \lambda\langle \vec{x}, \vec{x} \rangle \text{ since } \vec{x} \text{ has norm 1} \\
&= \langle \lambda\vec{x}, \vec{x} \rangle \text{ by linearity of } \langle \cdot, \cdot \rangle \\
&= \langle A\vec{x}, \vec{x} \rangle \text{ since } A\vec{x} = \lambda\vec{x} \\
&= (A\vec{x})^\top \vec{x} \text{ by definition of } \langle \cdot, \cdot \rangle \\
&= \vec{x}^\top \overline{(\bar{A}^\top \vec{x})} \text{ by expanding the product and using } \overline{ab} = \bar{a}\bar{b} \\
&= \langle \vec{x}, A^H\vec{x} \rangle \text{ by definition of } A^H \text{ and } \langle \cdot, \cdot \rangle \\
&= \langle \vec{x}, A\vec{x} \rangle \text{ since } A = A^H
\end{aligned}$$

$$= \bar{\lambda}\langle \vec{x}, \vec{x} \rangle \text{ since } A\vec{x} = \lambda\vec{x}$$
$$= \bar{\lambda} \text{ since } \vec{x} \text{ has norm 1}$$

Thus $\lambda = \bar{\lambda}$, which can happen only if $\lambda \in \mathbb{R}$, as needed. □

Not only are the eigenvalues of Hermitian (and symmetric) matrices real, but also their eigenvectors must be orthogonal:

Proposition 5.4. Eigenvectors corresponding to distinct eigenvalues of Hermitian matrices must be orthogonal.

Proof. Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian, and suppose $\lambda \neq \mu$ with $A\vec{x} = \lambda\vec{x}$ and $A\vec{y} = \mu\vec{y}$. By the previous proposition we know $\lambda, \mu \in \mathbb{R}$. Then, $\langle A\vec{x}, \vec{y} \rangle = \lambda\langle \vec{x}, \vec{y} \rangle$. But since A is Hermitian we can also write $\langle A\vec{x}, \vec{y} \rangle = \langle \vec{x}, A^H\vec{y} \rangle = \langle \vec{x}, A\vec{y} \rangle = \mu\langle \vec{x}, \vec{y} \rangle$. Thus, $\lambda\langle \vec{x}, \vec{y} \rangle = \mu\langle \vec{x}, \vec{y} \rangle$. Since $\lambda \neq \mu$, we must have $\langle \vec{x}, \vec{y} \rangle = 0$. □

Finally, we state without proof a crowning result of linear algebra, the Spectral Theorem. This theorem states that all symmetric or Hermitian matrices are non-defective and thus must have exactly n (orthogonal) eigenvectors.

Theorem 5.1 (Spectral Theorem)**.** Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian (if $A \in \mathbb{R}^{n\times n}$, suppose it is symmetric). Then, A has exactly n orthonormal eigenvectors $\vec{x}_1, \cdots, \vec{x}_n$ with (possibly repeated) eigenvalues $\lambda_1, \ldots, \lambda_n$. In other words, there exists an orthogonal matrix X of eigenvectors and diagonal matrix D of eigenvalues such that $D = X^\top AX$.

This theorem implies that any $\vec{y} \in \mathbb{R}^n$ can be decomposed into a linear combination of the eigenvectors of a Hermitian A. Many calculations are easier in this basis, as shown below:

Example 5.2 (Computation using eigenvectors)**.** Take $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^n$ to be the unit-length eigenvectors of a symmetric invertible matrix $A \in \mathbb{R}^{n\times n}$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. Suppose we wish to solve $A\vec{y} = \vec{b}$. By the Spectral Theorem, we can decompose $\vec{b} = c_1\vec{x}_1 + \cdots + c_n\vec{x}_n$, where $c_i = \vec{b} \cdot \vec{x}_i$ by orthonormality. Then,

$$\vec{y} = \frac{c_1}{\lambda_1}\vec{x}_1 + \cdots + \frac{c_n}{\lambda_n}\vec{x}_n.$$

The fastest way to check this formula is to multiply $\vec{y}$ by A and make sure we recover $\vec{b}$:

$$\begin{aligned} A\vec{y} &= A\left(\frac{c_1}{\lambda_1}\vec{x}_1 + \cdots + \frac{c_n}{\lambda_n}\vec{x}_n\right) \\ &= \frac{c_1}{\lambda_1}A\vec{x}_1 + \cdots + \frac{c_n}{\lambda_n}A\vec{x}_n \\ &= c_1\vec{x}_1 + \cdots + c_n\vec{x}_n \text{ since } A\vec{x}_k = \lambda_k\vec{x}_k \text{ for all } k \\ &= \vec{b}, \text{ as desired.} \end{aligned}$$

The calculation above is both a positive and negative result. It shows that given the eigenvectors and eigenvalues of symmetric matrix A, operations like inversion become straightforward. On the flip side, this means that finding the full set of eigenvectors of a symmetric matrix A is "at least" as difficult as solving $A\vec{x} = \vec{b}$.

Returning from our foray into the complex numbers, we revisit to real numbers to prove one final useful fact about positive definite matrices:

Proposition 5.5. All eigenvalues of positive definite matrices are nonnegative.

Proof. Take $A \in \mathbb{R}^{n\times n}$ positive definite, and suppose $A\vec{x} = \lambda\vec{x}$ with $\|\vec{x}\| = 1$. By positive definiteness, we know $\vec{x}^\top A\vec{x} \geq 0$. But, $\vec{x}^\top A\vec{x} = \vec{x}^\top(\lambda\vec{x}) = \lambda\|\vec{x}\|^2 = \lambda$, as needed. □

This property is not nearly as remarkable as those associated with symmetric or Hermitian matrices, but it helps *order* the eigenvalues of A. Positive definite matrices enjoy the property that the eigenvalue with smallest absolute value is also the eigenvalue closest to zero, and the eigenvalue with largest absolute value is the one farthest from zero. This property influences methods that seek only a subset of the eigenvalues of a matrix, usually at one of the two ends of its spectrum.

5.2.2 Specialized Properties

We mention some specialized properties of eigenvectors and eigenvalues that influence more advanced methods for their computation. They largely will not figure into our subsequent discussion, so this section can be skipped if readers lack sufficient background.

5.2.2.1 Characteristic Polynomial

The determinant of a matrix $\det A$ satisfies $\det A \neq 0$ if and only if A is invertible. Thus, one way to find eigenvalues of a matrix is to find roots of the *characteristic polynomial*

$$p_A(\lambda) = \det(A - \lambda I_{n\times n}).$$

We have chosen to avoid determinants in our discussion of linear algebra, but simplifying p_A reveals that it is an n-th degree polynomial in λ.

From this construction, we can define the *algebraic multiplicity* of an eigenvalue as its multiplicity as a root of p_A. The algebraic multiplicity of any eigenvalue is at least as large as the geometric multiplicity. If the algebraic multiplicity is 1, the root is called *simple*, because it corresponds to a single eigenvector that is linearly dependent with any others. Eigenvalues for which the algebraic and geometric multiplicities are not equal are called *defective*, since the corresponding matrix must also be defective in the sense of Definition 5.3.

In numerical analysis, it is common to avoid making use of the determinant of a matrix. While it is a convenient theoretical construction, its practical use is limited. Determinants are difficult to compute. In fact, eigenvalue algorithms do not attempt to find roots of p_A since doing so would require evaluation of a determinant. Furthermore, the determinant $\det A$ has *nothing* to do with the conditioning of A, so near-but-not-exactly zero determinant of $\det(A - \lambda I_{n\times n})$ might not show that λ is nearly an eigenvalue of A.

5.2.2.2 Jordan Normal Form

We can only diagonalize a matrix when it has a full eigenspace. All matrices, however, are similar to a matrix in Jordan normal form, a general layout satisfying the following criteria:

- Nonzero values are on the diagonal entries a_{ii} and on the "superdiagonal" $a_{i(i+1)}$.
- Diagonal values are eigenvalues repeated as many times as their algebraic multiplicity; the matrix is block diagonal about these clusters.
- Off-diagonal values are 1 or 0.

Thus, the shape looks something like the following

$$\begin{pmatrix} \lambda_1 & 1 & & & & & & \\ & \lambda_1 & 1 & & & & & \\ & & \lambda_1 & & & & & \\ & & & \lambda_2 & 1 & & & \\ & & & & \lambda_2 & & & \\ & & & & & \lambda_3 & & \\ & & & & & & \ddots & \end{pmatrix}$$

Jordan normal form is attractive theoretically because it always exists, but the $1/0$ structure is discrete and unstable under numerical perturbation.

5.3 COMPUTING A SINGLE EIGENVALUE

The computation and estimation of the eigenvalues of a matrix is a well-studied problem with many potential solutions. Each solution is tuned for a different situation, and achieving near-optimal conditioning or speed requires experimentation with several techniques. Here, we cover a few popular approaches to the eigenvalue problem encountered in practice.

5.3.1 Power Iteration

Assume that $A \in \mathbb{R}^{n \times n}$ is non-defective and nonzero, e.g. A is symmetric. Then, by definition, A has a full set of eigenvectors $\vec{x}_1, \ldots, \vec{x}_n \in \mathbb{R}^n$; we sort them such that their corresponding eigenvalues satisfy $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$.

Suppose we take an arbitrary vector $\vec{v}$. Since the eigenvectors of A span $\mathbb{R}^n$, we can write $\vec{v}$ in the $\vec{x}_i$ basis as $\vec{v} = c_1\vec{x}_1 + \cdots + c_n\vec{x}_n$. Applying A to both sides,

$$\begin{aligned} A\vec{v} &= c_1 A\vec{x}_1 + \cdots + c_n A\vec{x}_n \\ &= c_1\lambda_1\vec{x}_1 + \cdots + c_n\lambda_n\vec{x}_n \text{ since } A\vec{x}_i = \lambda_i\vec{x}_i \\ &= \lambda_1\left(c_1\vec{x}_1 + \frac{\lambda_2}{\lambda_1}c_2\vec{x}_2 + \cdots + \frac{\lambda_n}{\lambda_1}c_n\vec{x}_n\right) \\ A^2\vec{v} &= \lambda_1^2\left(c_1\vec{x}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^2 c_2\vec{x}_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^2 c_n\vec{x}_n\right) \\ &\vdots \\ A^k\vec{v} &= \lambda_1^k\left(c_1\vec{x}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^k c_2\vec{x}_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k c_n\vec{x}_n\right) \end{aligned}$$

As $k \to \infty$, the ratio $(\lambda_i/\lambda_1)^k \to 0$ unless $\lambda_i = \pm\lambda_1$, since λ_1 has the largest magnitude of any eigenvalue by construction. So, if $\vec{x}$ is the projection of $\vec{v}$ onto the space of eigenvectors with eigenvalues λ_1, then as $k \to \infty$ the following approximation begins to dominate: $A^k\vec{v} \approx \lambda_1^k\vec{x}$.

This argument leads to an exceedingly simple algorithm for computing a single eigenvector $\vec{x}_1$ of A corresponding to its largest-magnitude eigenvalue λ_1:

1. Take $\vec{v}_1 \in \mathbb{R}^n$ to be an arbitrary nonzero vector.
2. Iterate until convergence for increasing k: $\vec{v}_k = A\vec{v}_{k-1}$

```
function POWER-ITERATION(A)
    v ← ARBITRARY(n)
    for k ← 1, 2, 3, ...
        v ← Av
    return v
```

(a)

```
function NORMALIZED-ITERATION(A)
    v ← ARBITRARY(n)
    for k ← 1, 2, 3, ...
        w ← Av
        v ← w/‖w‖
    return v
```

(b)

FIGURE 5.3 Power iteration without (a) and with (b) normalization for finding the largest eigenvalue of a matrix.

```
function INVERSE-ITERATION(A)
    v ← ARBITRARY(n)
    for k ← 1, 2, 3, ...
        w ← A^{-1}v
        v ← w/‖w‖
    return v
```

(a)

```
function INVERSE-ITERATION-LU(A)
    v ← ARBITRARY(n)
    L, U ← LU-FACTORIZE(A)
    for k ← 1, 2, 3, ...
        y ← FORWARD-SUBSTITUTE(L, v)
        w ← BACK-SUBSTITUTE(U, y)
        v ← w/‖w‖
    return v
```

(b)

FIGURE 5.4 Inverse iteration without (a) and with (b) LU factorization.

This algorithm, known as *power iteration* and detailed in Figure 5.3(a), produces vectors $\vec{v}_k$ more and more parallel to the desired $\vec{x}_1$ as $k \to \infty$. Although we have not considered the defective case here, it is still guaranteed to converge; see [93] for a more advanced discussion.

One time that this technique may fail is if we accidentally choose $\vec{v}_1$ such that $c_1 = 0$, but the odds of this peculiarity occurring are vanishingly small. Such a failure mode only occurs when our initial guess has no component parallel to $\vec{x}_1$. Also, while power iteration can succeed in the presence of repeated eigenvalues, it can fail if both λ and $-\lambda$ are eigenvalues of A for some $\lambda \neq 0$. In the absence of these degeneracies, the rate of convergence for power iteration depends on the decay rate of terms 2 to n in the sum above for $A^k\vec{v}$ and hence is determined by the ratio of the second-largest-magnitude eigenvalue of A to the largest.

If $|\lambda_1| > 1$, however, then $\|\vec{v}_k\| \to \infty$ as $k \to \infty$, an undesirable property for floating point arithmetic. We only care about the *direction* of the eigenvector rather than its magnitude, so scaling has no effect on the quality of our solution. Thus, to avoid dealing with large-magnitude vectors we can normalize $\vec{v}_k$ at each step, producing the *normalized power iteration* algorithm in Figure 5.3(b). In the algorithm listing, we purposely do not decorate the norm $\|\cdot\|$ with a particular subscript. Mathematically, *any* norm will suffice for preventing $\vec{v}_k$ from going to infinity, since we have shown that all norms on $\mathbb{R}^n$ are equivalent. In practice, we often use the *infinity* norm $\|\cdot\|_\infty$; this choice has the convenient property that during iteration $\|A\vec{v}_k\|_\infty \to |\lambda_1|$.

5.3.2 Inverse Iteration

We now have an iteration for finding the *largest*-magnitude eigenvalue λ_1 of a matrix A. Suppose A is invertible, so that we can evaluate $\vec{y} = A^{-1}\vec{v}$ by solving $A\vec{y} = \vec{v}$ using

```
function RAYLEIGH-QUOTIENT-ITERATION(A, σ)
  v ← ARBITRARY(n)
  for k ← 1, 2, 3, ...
    σ ← (v^T A v) / ||v||_2
    w ← (A − σ I_{n×n})^{-1} v
    v ← w / ||w||
  return v
```

FIGURE 5.5 Rayleigh quotient iteration for finding an eigenvalue close to an initial guess σ.

techniques covered in previous chapters. If $A\vec{x} = \lambda\vec{x}$, then $\vec{x} = \lambda A^{-1}\vec{x}$, or equivalently $A^{-1}\vec{x} = \frac{1}{\lambda}\vec{x}$. Thus, $1/\lambda$ is an eigenvalue of A^{-1} with eigenvector $\vec{x}$.

If $|a| \geq |b|$ then $|b|^{-1} \geq |a|^{-1}$, so the smallest-magnitude eigenvalue of A is the *largest*-magnitude eigenvector of A^{-1}. This construction yields an algorithm for finding λ_n rather than λ_1 called *inverse power iteration*, as in Figure 5.4(a). This iterative scheme is nothing more than the power iteration method from §5.3.1 applied to A^{-1}.

We repeatedly are solving systems of equations using the same matrix A but different right hand sides, a perfect application of factorization techniques from previous chapters. For instance, if we write $A = LU$, then we could formulate an equivalent but considerably more efficient version of inverse power iteration illustrated in Figure 5.4(b). With this simplification, each solve for $A^{-1}\vec{v}$ is carried out in two steps, first by solving $L\vec{y} = \vec{v}$ and then by solving $U\vec{w} = \vec{y}$ as suggested in §2.5.1.

5.3.3 Shifting

Suppose λ_2 is the eigenvalue with second-largest magnitude of A. Power iteration converges fastest when $|\lambda_2/\lambda_1|$ is small, since in this case the power $(\lambda_2/\lambda_1)^k$ decays quickly. If this ratio is nearly 1, it may take many iterations before a single eigenvector is isolated.

If the eigenvalues of A are $\lambda_1, \ldots, \lambda_n$ with corresponding eigenvectors $\vec{x}_1, \ldots, \vec{x}_n$, then the eigenvalues of $A - \sigma I_{n\times n}$ are $\lambda_1 - \sigma, \ldots, \lambda_n - \sigma$, since:

$$(A - \sigma I_{n\times n})\vec{x}_i = A\vec{x}_i - \sigma\vec{x}_i = \lambda_i\vec{x}_i - \sigma\vec{x}_i = (\lambda_i - \sigma)\vec{x}_i.$$

With this idea in mind, one way to make power iteration converge quickly is to choose σ such that:

$$\left|\frac{\lambda_2 - \sigma}{\lambda_1 - \sigma}\right| < \left|\frac{\lambda_2}{\lambda_1}\right|.$$

That is, we find eigenvectors of $A - \sigma I_{n\times n}$ rather than A itself, choosing σ to widen the gap between the first and second eigenvalue to improve convergence rates. Guessing a good σ, however, can be an art, since we do not know the eigenvalues of A *a priori*.

More generally, if we think that σ is near an eigenvalue of A, then $A - \sigma I_{n\times n}$ has an eigenvalue close to 0 that we can reveal by inverse iteration. In other words, to use power iteration to target a particular eigenvalue of A rather than its largest or smallest eigenvalue as in previous sections, we shift A so that the eigenvalue we want is close to zero and then can apply inverse iteration to the result.

If our initial guess of σ is inaccurate, we could try to update it from iteration to iteration of the power method. For example, if we have a fixed guess of an eigenvector $\vec{x}$ of A, then

by the normal equations the least-squares approximation of the corresponding eigenvalue σ is given by

$$\sigma \approx \frac{\vec{x}^\top A\vec{x}}{\|\vec{x}\|_2^2}.$$

This fraction is known as a Rayleigh quotient. Thus, we can attempt to increase convergence by using *Rayleigh quotient iteration* as in Figure 5.5, which uses this approximation for σ to update the shift in each step.

Rayleigh quotient iteration usually takes fewer iterations to converge than power iteration given a good starting guess σ, but the matrix $A - \sigma_k I_{n\times n}$ is different each iteration and cannot be prefactored as in Figure 5.4(b). In other words, fewer iterations are necessary but each iteration takes more time; this trade-off makes the Rayleigh method more or less preferable to power iteration with a fixed shift depending on the particular choice and size of A. As an additional caveat, if σ_k is *too good* an estimate of an eigenvalue, the matrix $A - \sigma_k I_{n\times n}$ can become near-singular, causing conditioning issues during inverse iteration. In the opposite case, it can be difficult to control which eigenvalue is isolated by Rayleigh quotient iteration, especially if the initial guess is inaccurate.

5.4 FINDING MULTIPLE EIGENVALUES

So far, we have described techniques for finding a single eigenvalue/eigenvector pair: power iteration to find the largest eigenvalue, inverse iteration to find the smallest, and shifting to target values in between. For many applications, however, a single eigenvalue will not suffice. Thankfully, we can modify these techniques to handle this case as well.

5.4.1 Deflation

Recall the high-level structure of power iteration: Choose an arbitrary $\vec{v}_1$, and iteratively multiply it by A until only the largest eigenvalue λ_1 survives. Take $\vec{x}_1$ to be the corresponding eigenvector.

We were quick to dismiss an unlikely failure mode of this algorithm, however, when $\vec{v}_1 \cdot \vec{x}_1 = 0$, that is, when the initial eigenvector guess has no component parallel to $\vec{x}_1$. In this case, no matter how many times we apply A, the result will never have a component parallel to $\vec{x}_1$. The probability of choosing such a $\vec{v}_1$ is vanishingly small, so in all but the most pernicious of cases power iteration is a stable technique.

We can turn this drawback on its head to formulate a method for finding more than one eigenvalue of a symmetric matrix A. Suppose we find $\vec{x}_1$ and λ_1 via power iteration as before. After convergence, we can restart power iteration after projecting $\vec{x}_1$ out of the initial guess $\vec{v}_1$. Since the eigenvectors of A are orthogonal, by the argument in §5.3.1 power iteration after this projection will recover its *second*-largest eigenvalue!

Due to finite-precision arithmetic, applying A to a vector may inadvertently introduce a small component parallel to $\vec{x}_1$. We can avoid this effect by projecting in each iteration. This change yields the algorithm in Figure 5.6 for computing the eigenvalues in order of descending magnitude.

The inner loop of projected iteration is equivalent to power iteration on the matrix AP, where P projects out $\vec{v}_1, \ldots, \vec{v}_{\ell-1}$. AP has the same eigenvectors as A with eigenvalues $0, \ldots, 0, \lambda_\ell, \ldots, \lambda_n$. More generally, the method of *deflation* involves modifying the matrix A so that power iteration reveals an eigenvector that has not already been computed. For instance, AP is a modification of A so that the large eigenvalues we already have computed are zeroed out.

```
function PROJECTED-ITERATION(symmetric A,k)
  for ℓ ← 1, 2, ..., k
    v_ℓ ← ARBITRARY(n)
    for k ← 1, 2, 3, ...
      u ← v − proj_{span{v_1,...,v_{ℓ−1}}} v
      w ← Au
      v ← w/‖w‖
  return v_1, ..., v_k
```

FIGURE 5.6 Projection for finding k eigenvectors of a symmetric matrix A with the largest eigenvalues. If $\vec{u} = \vec{0}$ at any point, the remaining eigenvalues of A are all zero.

Projection can fail if A is asymmetric. Other deflation formulas, however, can work in its place with similar efficiency. For instance, suppose $A\vec{x}_1 = \lambda_1\vec{x}_1$ with $\|\vec{x}_1\| = 1$. Take H to be the Householder matrix (see §4.5) such that $H\vec{x}_1 = \vec{e}_1$, the first standard basis vector. From our discussion in §5.2, similarity transforms do not affect the set of eigenvalues, so we safely can conjugate A by H without significant change in eigenvalue structure. Consider what happens when we multiply $HAH^\top$ by $\vec{e}_1$:

$$\begin{aligned} HAH^\top\vec{e}_1 &= HAH\vec{e}_1 \text{ since } H \text{ is symmetric} \\ &= HA\vec{x}_1 \text{ since } H\vec{x}_1 = \vec{e}_1 \text{ and } H^2 = I_{n\times n} \\ &= \lambda_1 H\vec{x}_1 \text{ since } A\vec{x}_1 = \lambda_1\vec{x}_1 \\ &= \lambda_1\vec{e}_1 \text{ by definition of } H \end{aligned}$$

Thus, the first column of $HAH^\top$ is $\lambda_1\vec{e}_1$, showing that $HAH^\top$ has the following structure [54]:

$$HAH^\top = \begin{pmatrix} \lambda_1 & \vec{b}^\top \\ \vec{0} & B \end{pmatrix}.$$

The matrix $B \in \mathbb{R}^{(n-1)\times(n-1)}$ has eigenvalues $\lambda_2, \ldots, \lambda_n$. Thus, another algorithm for deflation iteratively generates smaller and smaller B matrices, with each eigenvalue computed using power iteration.

5.4.2 QR Iteration

Deflation has the drawback that we must compute each eigenvector separately, which can be slow and can accumulate error if individual eigenvalues are not accurate. Our remaining algorithms attempt to find more than one eigenvector simultaneously.

Recall that similar matrices A and $B = T^{-1}AT$ have the same eigenvalues for any invertible T. Thus, an algorithm seeking the eigenvalues of A can apply similarity transformations to A with abandon in the same way that Gaussian elimination premultiplies by row operations. Applying T^{-1} may be difficult, however, since it would require inverting T, so to make such a strategy practical we seek T's whose inverses are known.

One of our motivators for the QR factorization in Chapter 4 was that the matrix Q is *orthogonal*, satisfying $Q^{-1} = Q^\top$. In this class, Q and Q^{-1} are equally straightforward to apply, making orthogonal matrices strong choices for similarity transformations. We already applied this observation in §5.4.1 when we deflated using Householder matrices.

```
function QR-Iteration(A ∈ ℝ^{n×n})
  for k ← 1, 2, 3, ...
    Q, R ← QR-Factorize(A)
    A ← RQ
  return diag(R)
```

FIGURE 5.7 QR iteration for finding all the eigenvalues of A in the symmetric, non-repeated eigenvalue case.

But if we do not know *any* eigenvectors of A, which orthogonal matrix Q should we choose? Ideally, Q should involve the structure of A while being straightforward to compute. It is less clear how to apply Householder matrices strategically to reveal multiple eigenvalues in parallel,* but we do know how to generate one orthogonal Q from A by factoring $A = QR$. Then, experimentally we might conjugate A by Q to find:

$$Q^{-1}AQ = Q^\top AQ = Q^\top(QR)Q = (Q^\top Q)RQ = RQ$$

Amazingly, conjugating $A = QR$ by the orthogonal matrix Q is identical to writing the product RQ!

This matrix $A_2 \equiv RQ$ is *not* equal to $A = QR$, but it has the same eigenvalues. Hence, we can factor $A_2 = Q_2R_2$ to get a new orthogonal matrix Q_2, and once again conjugate to define $A_3 \equiv R_2Q_2$. Repeating this process indefinitely generates a whole sequence of similar matrices $A, A_2, A_3, \ldots$ with the same eigenvalues. Curiously, for many choices of A, as $k \to \infty$ one can check experimentally that while iterating QR factorization in this manner, R_k becomes a diagonal matrix containing the eigenvalues of A.

Based on this elegant observation, in the 1950s multiple groups of European mathematicians hypothesized the same iterative algorithm for finding the eigenvalues of a matrix A, shown in Figure 5.7:

Repeatedly factorize $A = QR$ and replace A with RQ.

Take A_k to be A after the k-th iteration of this method; that is $A_1 = A = Q_1R_1$, $A_2 = R_1Q_1 = Q_2R_2$, $A_3 = R_2Q_2 = Q_3R_3$, and so on. Since they are related via conjugation by a sequence of Q matrices, the matrices A_k all have the same eigenvalues as A. So, our analysis must show (1) when we expect this technique to converge and (2) if and how the limit point reveals eigenvectors of A. We will answer these questions in reverse order, for the case when A is symmetric with no repeated eigenvalues up to sign; so, if $\lambda \neq 0$ is an eigenvalue of A, then $-\lambda$ is not an eigenvalue of A. More advanced analysis and application to asymmetric or defective matrices can be found in [46] and elsewhere.

We begin by proving a proposition that will help us characterize limit behavior of the QR iteration algorithm:

Proposition 5.6. Take $A, B \in \mathbb{R}^{n\times n}$. Suppose that the eigenvectors of A span $\mathbb{R}^n$ and have distinct eigenvalues. Then, $AB = BA$ if and only if A and B have the same set of eigenvectors (with possibly different eigenvalues).

Proof. Suppose A and B have the same eigenvectors $\vec{x}_1, \ldots, \vec{x}_n$ with eigenvalues $\lambda_1^A, \ldots, \lambda_n^A$

*More advanced techniques do exactly this!

The conditions of this proposition can be relaxed but are sufficient for the discussion at hand.

for A and eigenvalues $\lambda_1^B, \ldots, \lambda_n^B$ for B. Any $\vec{y} \in \mathbb{R}^n$ can be decomposed as $\vec{y} = \sum_i a_i \vec{x}_i$, so:

$$BA\vec{y} = BA\sum_i a_i\vec{x}_i = B\sum_i \lambda_i^A \vec{x}_i = \sum_i \lambda_i^A \lambda_i^B \vec{x}_i$$
$$AB\vec{y} = AB\sum_i a_i\vec{x}_i = A\sum_i \lambda_i^B \vec{x}_i = \sum_i \lambda_i^A \lambda_i^B \vec{x}_i$$

So, $AB\vec{y} = BA\vec{y}$ for all $\vec{y} \in \mathbb{R}^n$, or equivalently $AB = BA$.

Now, suppose $AB = BA$, and take $\vec{x}$ to be any eigenvector of A with $A\vec{x} = \lambda\vec{x}$. Then, $A(B\vec{x}) = (AB)\vec{x} = (BA)\vec{x} = B(A\vec{x}) = \lambda(B\vec{x})$. We have two cases:

- If $B\vec{x} \neq \vec{0}$, then $B\vec{x}$ is an eigenvector of A with eigenvalue λ. Since A has no repeated eigenvalues and $\vec{x}$ is also an eigenvector of A with eigenvalue λ, we must have $B\vec{x} = c\vec{x}$ for some $c \neq 0$. In other words, $\vec{x}$ is also an eigenvector of B with eigenvalue c.
- If $B\vec{x} = \vec{0}$, then $\vec{x}$ is an eigenvector of B with eigenvalue 0.

Hence, all of the eigenvectors of A are eigenvectors of B. Since the eigenvectors of A span $\mathbb{R}^n$, A and B have exactly the same set of eigenvectors. □

Returning to QR iteration, suppose $A_k \to A_\infty$ as $k \to \infty$. If we factor $A_\infty = Q_\infty R_\infty$, then since QR iteration converged

$$A_\infty = Q_\infty R_\infty = R_\infty Q_\infty.$$

By the conjugation property, $Q_\infty^\top A_\infty Q_\infty = R_\infty Q_\infty = A_\infty$, or equivalently $A_\infty Q_\infty = Q_\infty A_\infty$. Since A_∞ has a full set of distinct eigenvalues, by Proposition 5.6, Q_∞ has the same eigenvectors as A_∞. The eigenvalues of Q_∞ are ± 1 by orthogonality. Suppose $A_\infty\vec{x} = \lambda\vec{x}$. Then,

$$\lambda\vec{x} = A_\infty\vec{x} = Q_\infty R_\infty\vec{x} = R_\infty Q_\infty\vec{x} = \pm R_\infty\vec{x},$$

so $R_\infty\vec{x} = \pm\lambda\vec{x}$. Since R_∞ is upper triangular, we now know (exercise 5.3):

The eigenvalues of A_∞—and hence the eigenvalues of A—are up to sign the diagonal elements of R_∞.

We can remove the sign caveat by computing QR factorization using rotations rather than reflections.

The derivation above assumes that there exists A_∞ with $A_k \to A_\infty$ as $k \to \infty$. Although we have not shown it yet, QR iteration is a stable method guaranteed to converge in many situations, and even when it does not converge, the relevant eigenstructure of A often can be computed from R_k as $k \to \infty$ regardless. We will not derive exact convergence conditions here but will provide some intuition for why we might expect this method to converge, at least given our restrictions on A.

To help motivate when we expect QR iteration to converge and yield eigenvalues along the diagonal of R_∞, suppose the columns of A are given by $\vec{a}_1, \ldots, \vec{a}_n$, and consider the matrix A^k for large k. We can write:

$$A^k = A^{k-1} \cdot A = \begin{pmatrix} | & | & & | \\ A^{k-1}\vec{a}_1 & A^{k-1}\vec{a}_2 & \cdots & A^{k-1}\vec{a}_n \\ | & | & & | \end{pmatrix}$$

By our derivation of power iteration, the first column of A^k will become more and more

parallel to the eigenvector $\vec{x}_1$ of A with largest magnitude $|\lambda_1|$ as $k \to \infty$, since we took a vector $\vec{a}_1$ and multiplied it by A many times.

Now, applying our intuition from deflation, suppose we project $\vec{x}_1$, which is approximately parallel to the first column of A^k, out of the second column of A^k. By orthogonality of the eigenvectors of A, we equivalently could have projected $\vec{x}_1$ out of $\vec{a}_2$ *initially* and then applied A^{k-1}. For this reason, as in §5.4.1, thanks to the removal of $\vec{x}_1$ the result of either process must be nearly parallel to $\vec{x}_2$, the vector with the *second*-most dominant eigenvalue! Proceeding inductively, when A is symmetric and thus has a full set of orthogonal eigenvectors, factoring $A^k = QR$ yields a set of near-eigenvectors of A in the columns of Q, in order of decreasing eigenvalue magnitude, with the corresponding eigenvalues along the diagonal of R.

Multiplying to find A^k for large k approximately takes the condition number of A to the k-th power, so computing the QR decomposition of A^k explicitly is likely to lead to numerical problems. Since decomposing A^k would reveal the eigenvector structure of A, however, we use this fact to our advantage without paying numerically. To do so, we make the following observation about QR iteration:

$$\begin{aligned}
A &= Q_1 R_1 \text{ by definition of QR iteration} \\
A^2 &= (Q_1 R_1)(Q_1 R_1) \\
&= Q_1 (R_1 Q_1) R_1 \text{ by regrouping} \\
&= Q_1 Q_2 R_2 R_1 \text{ since } A_2 = R_1 Q_1 = Q_2 R_2 \\
&\vdots \\
A^k &= Q_1 Q_2 \cdots Q_k R_k R_{k-1} \cdots R_1 \text{ by induction.}
\end{aligned}$$

Grouping the Q_i variables and the R_i variables separately provides a QR factorization of A^k. In other words, we can use the Q_k's and R_k's constructed during each step of QR iteration to construct a factorization of A^k, and thus we expect the columns of the product $Q_1 \cdots Q_k$ to converge to the eigenvectors of A.

By a similar argument, we show a related fact about the iterates $A_1, A_2, \ldots$ from QR iteration. Since $A_k = Q_k R_k$, we substitute $R_k = Q_k^\top A_k$ inductively to show:

$$\begin{aligned}
A_1 &= A \\
A_2 &= R_1 Q_1 \text{ by our construction of QR iteration} \\
&= Q_1^\top A Q_1 \text{ since } R_1 = Q_1^\top A_1 \\
A_3 &= R_2 Q_2 \\
&= Q_2^\top A_2 Q_2 \\
&= Q_2^\top Q_1^\top A Q_1 Q_2 \text{ from the previous step} \\
&\vdots \\
A_{k+1} &= Q_k^\top \cdots Q_1^\top A Q_1 \cdots Q_k \text{ inductively} \\
&= (Q_1 \cdots Q_k)^\top A (Q_1 \cdots Q_k),
\end{aligned}$$

where A_k is the k-th matrix from QR iteration. Thus, A_{k+1} is the matrix A conjugated by the product $\bar{Q}_k \equiv Q_1 \cdots Q_k$. We argued earlier that the columns of $\bar{Q}_k$ converge to the eigenvectors of A. Thus, since conjugating by the matrix of eigenvectors yields a diagonal matrix of eigenvalues, we know $A_{k+1} = \bar{Q}_k^\top A \bar{Q}$ will have approximate eigenvalues of A along its diagonal as $k \to \infty$, at least when eigenvalues are not repeated.

In the case of symmetric matrices without repeated eigenvalues, we have shown that both A_k and R_k will converge unconditionally to diagonal matrices containing the eigenvalues of A, while the product of the Q_k's will converge to a matrix of the corresponding eigenvectors. This case is but one example of the power of QR iteration, which is applied to many problems in which more than a few eigenvectors are needed of a given matrix A.

In practice, a few simplifying steps are usually applied before commencing QR iteration. QR factorization of a full matrix is relatively expensive computationally, so each iteration of the algorithm as we have described it is costly for large matrices. One way to avoid this cost for symmetric A is first to *tridiagonalize* A, systematically conjugating it by orthogonal matrices until entries not on or immediately adjacent to the diagonal are zero; tridiagonalization can be carried out using Householder matrices in $O(n^3)$ time for $A \in \mathbb{R}^{n\times n}$ [20]. QR factorization of symmetric tridiagonal matrices is much more efficient than the general case [87].

5.4.3 Krylov Subspace Methods

Our justification for QR iteration involved analyzing the columns of A^k as $k \to \infty$ applying observations we already made about power iteration in §5.3.1. More generally, for a vector $\vec{b} \in \mathbb{R}^n$, we can examine the so-called *Krylov matrix*

$$K_k \equiv \begin{pmatrix} | & | & | & & | \\ \vec{b} & A\vec{b} & A^2\vec{b} & \cdots & A^{k-1}\vec{b} \\ | & | & | & & | \end{pmatrix}.$$

Methods analyzing K_k to find eigenvectors and eigenvalues generally are classified as *Krylov subspace methods*. For instance, the *Arnoldi iteration* algorithm uses Gram-Schmidt orthogonalization to maintain an orthogonal basis $\{\vec{q}_1, \ldots, \vec{q}_k\}$ for the column space of K_k:

1. Begin by taking $\vec{q}_1$ to be an arbitrary unit-norm vector
2. For $k = 2, 3, \ldots$
   (a) Take $\vec{a}_k = A\vec{q}_{k-1}$
   (b) Project out the $\vec{q}$'s you already have computed:
   $$\vec{b}_k = \vec{a}_k - \text{proj}_{\text{span}\,\{\vec{q}_1,\ldots,\vec{q}_{k-1}\}}\vec{a}_k$$
   (c) Renormalize to find the next $\vec{q}_k = \vec{b}_k/\|\vec{b}_k\|$.

The matrix Q_k whose columns are the vectors found above is an orthogonal matrix with the same column space as K_k, and eigenvalue estimates can be recovered from the structure of $Q_k^\top A Q_k$.

The use of Gram-Schmidt makes this technique unstable and timing gets progressively worse as k increases, so extensions are needed to make it feasible. For instance, one approach involves running some iterations of Arnoldi, using the output to generate a better guess for the initial $\vec{q}_1$, and restarting [75]. Methods in this class are suited for problems requiring multiple eigenvectors at one of the ends of the spectrum without computing the complete set. They also can be applied to designing iterative methods for solving linear systems of equations, as we will explore in Chapter 10.

5.5 SENSITIVITY AND CONDITIONING

We have only outlined a few eigenvalue techniques out of a rich and long-standing literature. Almost any algorithmic technique has been experimented with for finding spectra, from iterative methods to root-finding on the characteristic polynomial to methods that divide matrices into blocks for parallel processing.

As with linear solvers, we can evaluate the conditioning of an eigenvalue problem independently of the solution technique. This analysis can help understand whether a simplistic iterative scheme will be successful for finding the eigenvectors of a given matrix or if more complex stabilized methods are necessary. To do so, we will derive a condition number for the problem of finding eigenvalues for a given matrix A. Before proceeding, we should highlight that the conditioning of an eigenvalue problem is *not* the same as the condition number of the matrix for solving linear systems.

Suppose a matrix A has an eigenvector $\vec{x}$ with eigenvalue λ. Analyzing the conditioning of the eigenvalue problem involves analyzing the stability of $\vec{x}$ and λ to perturbations in A. To this end, we might perturb A by a small matrix δA, thus changing the set of eigenvectors. We can write eigenvectors of $A + \delta A$ as perturbations of eigenvectors of A by solving the problem

$$(A + \delta A)(\vec{x} + \delta\vec{x}) = (\lambda + \delta\lambda)(\vec{x} + \delta\vec{x}).$$

Expanding both sides yields:

$$A\vec{x} + A\delta\vec{x} + \delta A \cdot \vec{x} + \delta A \cdot \delta\vec{x} = \lambda\vec{x} + \lambda\delta\vec{x} + \delta\lambda \cdot \vec{x} + \delta\lambda \cdot \delta\vec{x}$$

Since δA is small, we will assume* that $\delta\vec{x}$ and $\delta\lambda$ also are small. Products between these variables then are negligible, yielding the following approximation:

$$A\vec{x} + A\delta\vec{x} + \delta A \cdot \vec{x} \approx \lambda\vec{x} + \lambda\delta\vec{x} + \delta\lambda \cdot \vec{x}$$

Since $A\vec{x} = \lambda\vec{x}$, we can subtract this vector from both sides to find:

$$A\delta\vec{x} + \delta A \cdot \vec{x} \approx \lambda\delta\vec{x} + \delta\lambda \cdot \vec{x}$$

We now apply an analytical trick to complete our derivation. Since $A\vec{x} = \lambda\vec{x}$, we know $(A - \lambda I_{n\times n})\vec{x} = \vec{0}$, so $A - \lambda I_{n\times n}$ is not full rank. The transpose of a matrix is full-rank only if the matrix is full-rank, so we know $(A - \lambda I_{n\times n})^\top = A^\top - \lambda I_{n\times n}$ also has a null space vector $\vec{y}$. Thus $A^\top\vec{y} = \lambda\vec{y}$; we can call $\vec{y}$ a *left* eigenvector corresponding to $\vec{x}$. Left-multiplying our perturbation estimate above by $\vec{y}^\top$ shows

$$\vec{y}^\top(A\delta\vec{x} + \delta A \cdot \vec{x}) \approx \vec{y}^\top(\lambda\delta\vec{x} + \delta\lambda \cdot \vec{x}).$$

Since $A^\top\vec{y} = \lambda\vec{y}$, we can simplify:

$$\vec{y}^\top\delta A \cdot \vec{x} \approx \delta\lambda\vec{y}^\top\vec{x}$$

Rearranging yields:

$$\delta\lambda \approx \frac{\vec{y}^\top(\delta A)\vec{x}}{\vec{y}^\top\vec{x}}$$

Finally, assume $\|\vec{x}\| = 1$ and $\|\vec{y}\| = 1$. Then, taking norms on both sides shows:

$$|\delta\lambda| \lessapprox \frac{\|\delta A\|_2}{|\vec{y} \cdot \vec{x}|}$$

*This assumption should be checked in a more rigorous treatment!

So, conditioning of the eigenvalue problem depends directly on the size of the perturbation δA—as expected—and inversely on the angle between the left and right eigenvectors $\vec{x}$ and $\vec{y}$.

Based on this derivation, we can use $1/\vec{x}\cdot\vec{y}$ as an approximate condition number for finding the eigenvector $\vec{x}$ of A. Symmetric matrices have the same left and right eigenvectors, so $\vec{x} = \vec{y}$, yielding a condition number of 1. This strong conditioning reflects the fact that the eigenvectors of symmetric matrices are orthogonal and thus maximally separated.

5.6 EXERCISES

5.1 Verify the solution $\vec{x}(t)$ given in §5.1.2 to the ODE $\vec{x}' = A\vec{x}$.

5.2 Define

$$A \equiv \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Can power iteration find eigenvalues of this matrix? Why or why not?

5.3 Show that the eigenvalues of upper triangular matrices $U \in \mathbb{R}^{n\times n}$ are exactly their diagonal elements.

5.4 Extending problem 5.3, if we assume that the eigenvectors of U are $\vec{v}_k$ satisfying $U\vec{v}_k = u_{kk}\vec{v}_k$, characterize $\text{span}\{\vec{v}_1, \ldots, \vec{v}_k\}$ for $1 \leq k \leq n$ when the diagonal values u_{kk} of U are distinct.

5.5 We showed that the Rayleigh quotient iteration method can converge more quickly than power iteration. Why, however, might it still be more efficient to use the power method in some cases?

5.6 Suppose $\vec{u}$ and $\vec{v}$ are vectors in $\mathbb{R}^n$ such that $\vec{u}^\top\vec{v} = 1$, and define $A \equiv \vec{u}\vec{v}^\top$.

(a) What are the eigenvalues of A?

(b) How many iterations does power iteration take to converge to the dominant eigenvalue of A?

5.7 Suppose $B \in \mathbb{R}^{n\times n}$ is diagonalizable with eigenvalues λ_i satisfying $0 < \lambda_1 = \lambda_2 < \lambda_3 < \cdots < \lambda_n$. Let $\vec{v}_i$ be the eigenvector corresponding to λ_i. Show that the inverse power method applied to B converges to a linear combination of $\vec{v}_1$ and $\vec{v}_2$.

5.8 ("Mini-Riesz Representation Theorem") We will say $\langle\cdot,\cdot\rangle$ is an *inner product* on $\mathbb{R}^n$ if it satisfies:

a. $\langle\vec{x},\vec{y}\rangle = \langle\vec{y},\vec{x}\rangle\ \forall\vec{x},\vec{y} \in \mathbb{R}^n$

b. $\langle\alpha\vec{x},\vec{y}\rangle = \alpha\langle\vec{x},\vec{y}\rangle\ \forall\vec{x},\vec{y} \in \mathbb{R}^n, \alpha \in \mathbb{R}$

c. $\langle\vec{x}+\vec{y},\vec{z}\rangle = \langle\vec{x},\vec{z}\rangle + \langle\vec{y},\vec{z}\rangle\ \forall\vec{x},\vec{y},\vec{z} \in \mathbb{R}^n$

d. $\langle\vec{x},\vec{x}\rangle \geq 0$ with equality if and only if $\vec{x} = \vec{0}$.

Here we will derive a special case of a theorem applied in geometry processing and machine learning:

(a) Show that for a given inner product $\langle \cdot, \cdot \rangle$ there exists a corresponding matrix A such that $\langle \vec{x}, \vec{y} \rangle = \vec{x}^\top A \vec{y}$. For the same inner product, also show that there exists a matrix M such that $\langle \vec{x}, \vec{y} \rangle = (M\vec{x}) \cdot (M\vec{y})$. [This shows that all inner products are dot products after suitable rotation, stretching, and shearing of $\mathbb{R}^n$!]

(b) A *Mahalanobis metric* on $\mathbb{R}^n$ is a distance function of the form $d(\vec{x}, \vec{y}) = \sqrt{\langle \vec{x} - \vec{y}, \vec{x} - \vec{y} \rangle}$ for inner product $\langle \cdot, \cdot \rangle$. Use the result of 5.8a to give an alternative description of the set of Mahalanobis metrics.

(c) Suppose we are given a series of pairs $(\vec{x}_i, \vec{y}_i) \in \mathbb{R}^n \times \mathbb{R}^n$. A typical "metric learning" problem might involve finding a nontrivial Mahalanobis metric such that each $\vec{x}_i$ is close to each $\vec{y}_i$ with respect to that metric. Propose an optimization for this problem in the form of problem 0.7.
Note: Make sure that your optimal Mahalanobis distance is nonzero, but it is acceptable if your optimization allows *pseudometrics*, that is, there can exist $\vec{x} \neq \vec{y}$ with $d(\vec{x}, \vec{y}) = 0$. Solving this problem numerically will require machinery from Chapter 5; we include it here for practice designing optimization problems using linear algebra.

5.9 Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite.

(a) Define a matrix $\sqrt{A} \in \mathbb{R}^{n \times n}$ and show that $(\sqrt{A})^2 = A$. Generally speaking $\sqrt{A}$ is not the same as L in the Cholesky factorization $A = LL^\top$.

(b) Do most matrices have unique square roots? Why or why not?

(c) We can define the *exponential* of A as $e^A \equiv \sum_{k=0}^{\infty} \frac{1}{k!} A^k$; this sum is unconditionally convergent (you do not have to prove this!). Write an alternative expression for e^A in terms of the eigenvectors and eigenvalues of A.

(d) If $AB = BA$, show $e^{A+B} = e^A e^B$.

(e) Show that the ordinary differential equation $\vec{y}'(t) = -A\vec{y}$ with $\vec{y}(0) = \vec{y}_0$ for some $\vec{y}_0 \in \mathbb{R}^n$ is solved by $\vec{y}(t) = e^{-At}\vec{y}_0$. What happens as $t \to \infty$?

5.10 ("Epidemiology") Suppose $\vec{x}_0 \in \mathbb{R}^n$ contains sizes of different populations carrying a particular infection in year 0; for example, when tracking malaria we might take x_{01} to be the number of humans with malaria and x_{02} to be the number of mosquitoes carrying the disease. By writing relationships like "The average mosquito infects two humans" we can write a matrix M such that $\vec{x}_1 \equiv M\vec{x}_0$ predicts populations in year 1, $\vec{x}_2 \equiv M^2\vec{x}_0$ predicts populations in year 2, and so on.

(a) The spectral radius $\rho(M)$ is given by $\max_i |\lambda_i|$, where the eigenvalues of M are $\lambda_1, \ldots, \lambda_k$. Epidemiologists call this number the "reproduction number" $\mathcal{R}_0$ of M. Explain the difference between the cases $\mathcal{R}_0 < 1$ and $\mathcal{R}_0 > 1$ in terms of the spread of disease. Which case is more dangerous?

(b) Suppose we only care about proportions. For instance, we might use $M \in \mathbb{R}^{50 \times 50}$ to model transmission of diseases between residents in each of the 50 states of the USA, and we only care about the fraction of the total people with a disease who live in each state. If $\vec{y}_0$ holds these proportions in year 0, give an iterative scheme to predict proportions in future years. Characterize behavior as time goes to infinity.

Note: Those readers concerned about computer graphics applications of this material should know that the reproduction number $\mathcal{R}_0$ is referenced in the 2011 thriller *Contagion*.

5.11 ("Normalized cuts," [105]) Similar to spectral embedding (§5.1.3), suppose we have a collection of n objects and a symmetric matrix $W \in (\mathbb{R}^+)^{n\times n}$ whose entries w_{ij} measure the similarity between object i and object j. Rather than computing an embedding, however, now we would like to *cluster* the objects into two groups. This machinery is used to mark photos as day or night and to classify pixels in an image as foreground or background.

(a) Suppose we cluster $\{1,\dots,n\}$ into two disjoint sets A and B; this clustering defines a *cut* of the collection. We define the *cut score* of (A,B) as follows:

$$C(A,B) \equiv \sum_{\substack{i\in A\\ j\in B}} w_{ij}.$$

This score is large if objects in A and B are similar. Efficiency aside, why is it inadvisable to minimize $C(A,B)$ with respect to A and B?

(b) Define the *volume* of a set A as $V(A) \equiv \sum_{i\in A}\sum_{j=1}^{n} w_{ij}$. To alleviate issues with minimizing the cut score directly, instead we will attempt minimize the *normalized cut* score $N(A,B) \equiv C(A,B)(V(A)^{-1} + V(B)^{-1})$. What does this score measure?

(c) For a fixed choice of A and B, define $\vec{x} \in \mathbb{R}^n$ such that

$$x_i \equiv \begin{cases} V(A)^{-1} & \text{if } i \in A \\ -V(B)^{-1} & \text{if } i \in B. \end{cases}$$

Define matrices L and D such that

$$\vec{x}^\top L\vec{x} = \sum_{\substack{i\in A\\ j\in B}} w_{ij}\left(V(A)^{-1} + V(B)^{-1}\right)^2$$
$$\vec{x}^\top D\vec{x} = V(A)^{-1} + V(B)^{-1}.$$

Conclude that $N(A,B) = \frac{\vec{x}^\top L\vec{x}}{\vec{x}^\top D\vec{x}}$.

(d) Show that $\vec{x}^\top D\vec{1} = \vec{0}$.

(e) The *normalized cuts algorithm* computes A and B by optimizing for $\vec{x}$. Argue that the result of the following optimization lower-bounds the minimum normalized cut score of any partition (A,B) :

$$\begin{array}{rl} \min_{\vec{x}} & \frac{\vec{x}^\top L\vec{x}}{\vec{x}^\top D\vec{x}} \\ \text{such that} & \vec{x}^\top D\vec{1} = \vec{0}. \end{array}$$

Assuming D is invertible, show that this relaxed $\vec{x}$ can be computed using an eigenvalue problem.

CHAPTER 6

Singular Value Decomposition

CONTENTS

6.1 Deriving the SVD 125
6.1.1 Computing the SVD 127
6.2 Applications of the SVD 128
6.2.1 Solving Linear Systems and the Pseudoinverse 128
6.2.2 Decomposition into Outer Products and Low-Rank Approximations 129
6.2.3 Matrix Norms 130
6.2.4 The Procrustes Problem and Point Cloud Alignment 131
6.2.5 Principal Component Analysis (PCA) 133
6.2.6 Eigenfaces 133

IN Chapter 5, we derived a number of algorithms for computing the eigenvalues and eigenvectors of matrices $A \in \mathbb{R}^{n\times n}$. Having developed this machinery, we complete our initial discussion of numerical linear algebra by deriving and making use of one final matrix factorization that exists for *any* matrix $A \in \mathbb{R}^{m\times n}$, even if it is not symmetric or square: the singular value decomposition (SVD).

6.1 DERIVING THE SVD

For $A \in \mathbb{R}^{m\times n}$, we can think of the function $\vec{x} \mapsto A\vec{x}$ as a map taking points in $\mathbb{R}^n$ to points in $\mathbb{R}^m$. From this perspective, we might ask what happens to the geometry of $\mathbb{R}^n$ in the process, and in particular the effect A has on lengths of and angles between vectors.

Applying our usual starting point for eigenvalue problems, we examine the effect that A has on the lengths of vectors by examining critical points of the ratio

$$R(\vec{x}) = \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2}$$

over various vectors $\vec{x}$. This quotient measures relative shrinkage or growth of $\vec{x}$ under the action of A. Scaling $\vec{x}$ does not matter, since

$$R(\alpha\vec{x}) = \frac{\|A\cdot\alpha\vec{x}\|_2}{\|\alpha\vec{x}\|_2} = \frac{|\alpha|}{|\alpha|}\cdot\frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2} = \frac{\|A\vec{x}\|_2}{\|\vec{x}\|_2} = R(\vec{x}).$$

Thus, we can restrict our search to $\vec{x}$ with $\|\vec{x}\| = 1$. Furthermore, since $R(\vec{x}) \geq 0$, we can instead consider $[R(\vec{x})]^2 = \|A\vec{x}\|_2^2 = \vec{x}^\top A^\top A\vec{x}$. As we have shown in previous chapters,

critical points of $\vec{x}^\top A^\top A\vec{x}$ subject to $\|\vec{x}\|_2 = 1$ are exactly the eigenvectors $\vec{x}_i$ satisfying $A^\top A\vec{x}_i = \lambda_i\vec{x}_i$; we know $\lambda_i \geq 0$ and $\vec{x}_i \cdot \vec{x}_j = 0$ when $i \neq j$ since $A^\top A$ is symmetric and positive semidefinite.

Based on our use of the function R, the $\{\vec{x}_i\}$ basis is a reasonable one for studying the effects of A on $\mathbb{R}^n$. Returning to the original goal of characterizing the action of A from a geometric standpoint, define $\vec{y}_i \equiv A\vec{x}_i$. We can make an additional observation about $\vec{y}_i$ revealing a second eigenvalue structure:

$$\begin{aligned}
\lambda_i\vec{y}_i &= \lambda_i \cdot A\vec{x}_i \text{ by definition of } \vec{y}_i \\
&= A(\lambda_i\vec{x}_i) \\
&= A(A^\top A\vec{x}_i) \text{ since } \vec{x}_i \text{ is an eigenvector of } A^\top A \\
&= (AA^\top)(A\vec{x}_i) \text{ by associativity} \\
&= (AA^\top)\vec{y}_i
\end{aligned}$$

Thus, we have two cases:

1. Suppose $\vec{y}_i \neq \vec{0}$. In this case, $\vec{y}_i = A\vec{x}_i$ is a *corresponding* eigenvector of $AA^\top$ with $\|\vec{y}_i\|_2 = \|A\vec{x}_i\|_2 = \sqrt{\|A\vec{x}_i\|_2^2} = \sqrt{\vec{x}_i^\top A^\top A\vec{x}_i} = \sqrt{\lambda_i}\|\vec{x}_i\|_2$.
2. Otherwise, $\vec{y}_i = \vec{0}$.

An identical proof shows that if $\vec{y}$ is an eigenvector of $AA^\top$, then $\vec{x} \equiv A^\top\vec{y}$ is either zero or an eigenvector of $A^\top A$ with the same eigenvalue.

Take k to be the number of strictly positive eigenvalues $\lambda_i > 0$ for $i \in \{1, \ldots, k\}$. By our construction above, we can take $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$ to be eigenvectors of $A^\top A$ and corresponding eigenvectors $\vec{y}_1, \ldots, \vec{y}_k \in \mathbb{R}^m$ of $AA^\top$ such that

$$\begin{aligned}
A^\top A\vec{x}_i &= \lambda_i\vec{x}_i \\
AA^\top\vec{y}_i &= \lambda_i\vec{y}_i
\end{aligned}$$

for eigenvalues $\lambda_i > 0$; here we normalize such that $\|\vec{x}_i\|_2 = \|\vec{y}_i\|_2 = 1$ for all i. Following traditional notation, we can define matrices $\bar{V} \in \mathbb{R}^{n\times k}$ and $\bar{U} \in \mathbb{R}^{m\times k}$ whose columns are $\vec{x}_i$'s and $\vec{y}_i$'s, resp.

We can examine the effect of these new basis matrices on A. Take $\vec{e}_i$ to be the i-th standard basis vector. Then,

$$\begin{aligned}
\bar{U}^\top A\bar{V}\vec{e}_i &= \bar{U}^\top A\vec{x}_i \text{ by definition of } \bar{V} \\
&= \frac{1}{\lambda_i}\bar{U}^\top A(\lambda_i\vec{x}_i) \text{ since we assumed } \lambda_i > 0 \\
&= \frac{1}{\lambda_i}\bar{U}^\top A(A^\top A\vec{x}_i) \text{ since } \vec{x}_i \text{ is an eigenvector of } A^\top A \\
&= \frac{1}{\lambda_i}\bar{U}^\top (AA^\top)A\vec{x}_i \text{ by associativity} \\
&= \frac{1}{\sqrt{\lambda_i}}\bar{U}^\top (AA^\top)\vec{y}_i \text{ since we rescaled so that } \|\vec{y}_i\|_2 = 1 \\
&= \sqrt{\lambda_i}\bar{U}^\top\vec{y}_i \text{ since } AA^\top\vec{y}_i = \lambda_i\vec{y}_i \\
&= \sqrt{\lambda_i}\vec{e}_i
\end{aligned}$$

Take $\bar{\Sigma} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k})$. Then, the derivation above shows that $\bar{U}^\top A\bar{V} = \bar{\Sigma}$.

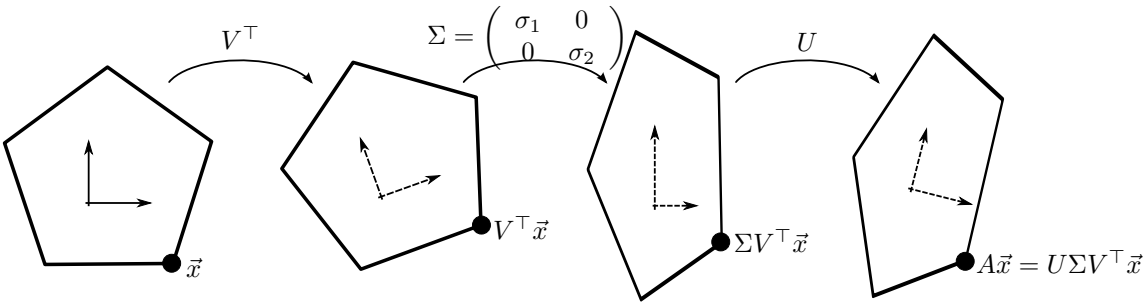


FIGURE 6.1 Geometric interpretation for the singular value decomposition $A = U\Sigma V^\top$. The matrices U and $V^\top$ are orthogonal and hence can only rotate or reflect vectors without changing their length. The diagonal matrix Σ scales the horizontal and vertical axes.

Complete the columns of $\bar{U}$ and $\bar{V}$ to $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ by adding orthonormal null space vectors $\vec{x}_i$ and $\vec{y}_i$ with $A^\top A\vec{x}_i = \vec{0}$ and $AA^\top\vec{y}_i = \vec{0}$, resp. After this extension, $U^\top AV\vec{e}_i = \vec{0}$ and/or $\vec{e}_i^\top U^\top AV = \vec{0}^\top$ for $i > k$. Thus, if we take

$$\Sigma_{ij} \equiv \begin{cases} \sqrt{\lambda_i} & i = j \text{ and } i \leq k \\ 0 & \text{otherwise} \end{cases}$$

then we can extend our previous relationship to show $U^\top AV = \Sigma$, or by orthogonality of U and V,

$$A = U\Sigma V^\top.$$

This factorization is the *singular value decomposition* (SVD) of A. The columns of U are called the *left singular vectors*, and the columns of V are called the *right singular vectors*. The diagonal elements σ_i of Σ are the *singular values* of A; usually they are sorted such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Both U and V are orthogonal matrices; the columns of U and V corresponding to $\sigma_i \neq 0$ span the column and row spaces of A, resp.

The SVD provides a complete geometric characterization of the action of A. Since U and V are orthogonal, they can be thought of as rotation matrices; as a diagonal matrix, Σ scales individual coordinates. Thus, *all* matrices $A \in \mathbb{R}^{m\times n}$ are a composition of a rotation, a scale, and a second rotation. This sequence of operations is illustrated in Figure 6.1.

6.1.1 Computing the SVD

The columns of V are the eigenvectors of $A^\top A$, so they can be computed using techniques discussed in the previous chapter. Since $A = U\Sigma V^\top$, we know $AV = U\Sigma$. Thus, the columns of U corresponding to nonzero singular values in Σ are normalized columns of AV; the remaining columns satisfy $AA^\top\vec{u}_i = \vec{0}$, which can be solved using LU factorization.

This is by no means the most efficient or stable way to compute the SVD, but it works reasonably well for many applications. We will omit more specialized algorithms for finding the SVD, but many of them are extensions of power iteration and other algorithms we already have covered that operate without forming $A^\top A$ or $AA^\top$ explicitly.

6.2 APPLICATIONS OF THE SVD

We devote the remainder of this chapter introducing applications of the SVD. The SVD appears countless times in both the theory and practice of numerical linear algebra, and its importance hardly can be exaggerated.

6.2.1 Solving Linear Systems and the Pseudoinverse

In the special case where $A \in \mathbb{R}^{n \times n}$ is square and invertible, the SVD can be used to solve the linear problem $A\vec{x} = \vec{b}$. In particular, we have $U\Sigma V^\top \vec{x} = \vec{b}$, or

$$\vec{x} = V\Sigma^{-1}U^\top \vec{b}.$$

In this case Σ is a square diagonal matrix, so Σ^{-1} is the matrix whose diagonal entries are $1/\sigma_i$.

Computing the SVD is far more expensive than most of the linear solution techniques we introduced in Chapter 2, so this initial observation mostly is of theoretical rather than practical interest. More generally, however, suppose we wish to find a least-squares solution to $A\vec{x} \approx \vec{b}$, where $A \in \mathbb{R}^{m \times n}$ is not necessarily square. From our discussion of the normal equations, we know that $\vec{x}$ must satisfy $A^\top A\vec{x} = A^\top \vec{b}$. But when A is "short" or "underdetermined," that is, when A has more columns than rows ($m < n$) or has linearly dependent columns, the solution to the normal equations might not be unique.

To cover the under-, completely-, and over-determined cases simultaneously, we can solve an optimization problem of the following form:

$$\begin{array}{ll} \text{minimize} & \|\vec{x}\|_2^2 \\ \text{such that} & A^\top A\vec{x} = A^\top \vec{b} \end{array}$$

This optimization asks that $\vec{x}$ satisfy the normal equations with the least possible norm. When $A^\top A$ is invertible, there is only one $\vec{x}$ that satisfies the constraints. Otherwise, of all the feasible vectors $\vec{x}$ we choose the one with smallest $\|\vec{x}\|_2$; that is, we seek the "simplest possible" least-square solution of $A\vec{x} \approx \vec{b}$, when multiple $\vec{x}$'s minimize $\|A\vec{x} - \vec{b}\|_2$.

Write $A = U\Sigma V^\top$. Then,

$$\begin{aligned} A^\top A &= (U\Sigma V^\top)^\top (U\Sigma V^\top) \\ &= V\Sigma^\top U^\top U\Sigma V^\top \text{ since } (AB)^\top = B^\top A^\top \\ &= V\Sigma^\top \Sigma V^\top \text{ since } U \text{ is orthogonal} \end{aligned}$$

Thus, asking that $A^\top A\vec{x} = A^\top \vec{b}$ is the same as writing

$$V\Sigma^\top \Sigma V^\top \vec{x} = V\Sigma U^\top \vec{b}$$

$$\text{Or equivalently, } \Sigma\vec{y} = \vec{d}$$

if we take $\vec{d} \equiv U^\top \vec{b}$ and $\vec{y} \equiv V^\top \vec{x}$. Since U is orthogonal, $\|\vec{y}\|_2 = \|\vec{x}\|_2$ and our optimization becomes:

$$\begin{array}{ll} \text{minimize} & \|\vec{y}\|_2^2 \\ \text{such that} & \Sigma\vec{y} = \vec{d} \end{array}$$

Since Σ is diagonal, however, the condition $\Sigma\vec{y} = \vec{d}$ can be written $\sigma_i y_i = d_i$. So, whenever $\sigma_i \neq 0$ we must have $y_i = d_i/\sigma_i$. When $\sigma_i = 0$, there is *no* constraint on y_i, so since we

are minimizing $\|\vec{y}\|_2^2$ we might as well take $y_i = 0$. In other words, the solution to this optimization is $\vec{y} = \Sigma^+ \vec{d}$, where $\Sigma^+ \in \mathbb{R}^{n \times m}$ has the following form:

$$\Sigma^+_{ij} \equiv \begin{cases} 1/\sigma_i & i = j \text{ and } \sigma_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Undoing our change of variables, this result in turn yields $\vec{x} = V\vec{y} = V\Sigma^+\vec{d} = V\Sigma^+U^\top\vec{b}$.

With this motivation, we make the following definition:

Definition 6.1 (Pseudoinverse)**.** The *pseudoinverse* of $A = U\Sigma V^\top \in \mathbb{R}^{m \times n}$ is $A^+ \equiv V\Sigma^+U^\top \in \mathbb{R}^{n \times m}$.

Our derivation above shows that the pseudoinverse of A enjoys the following properties:

- When A is square and invertible, $A^+ = A^{-1}$.
- When A is overdetermined, $A^+\vec{b}$ gives the least-squares solution to $A\vec{x} \approx \vec{b}$.
- When A is underdetermined, $A^+\vec{b}$ gives the least-squares solution to $A\vec{x} \approx \vec{b}$ with minimal (Euclidean) norm.

In this way, the SVD allows us to unify solutions of the underdetermined, fully-determined, and overdetermined cases of $A\vec{x} \approx \vec{b}$.

6.2.2 Decomposition into Outer Products and Low-Rank Approximations

If we expand the product $A = U\Sigma V^\top$ column-by-column, an equivalent formula is the following:

$$A = \sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top,$$

where $\ell \equiv \min\{m, n\}$, and $\vec{u}_i$ and $\vec{v}_i$ are the i-th columns of U and V, resp. Our sum only goes to $\min\{m, n\}$ since we know that the remaining columns of U or V will be zeroed out by Σ.

This expression shows that any matrix can be decomposed as the sum of *outer products* of vectors:

Definition 6.2 (Outer product)**.** The *outer product* of $\vec{u} \in \mathbb{R}^m$ and $\vec{v} \in \mathbb{R}^n$ is the matrix $\vec{u} \otimes \vec{v} \equiv \vec{u}\vec{v}^\top \in \mathbb{R}^{m \times n}$.

This alternative formula for the SVD provides a new way to find the product $A\vec{x}$:

$$\begin{aligned} A\vec{x} &= \left(\sum_{i=1}^{\ell} \sigma_i \vec{u}_i \vec{v}_i^\top \right) \vec{x} \\ &= \sum_{i=1}^{\ell} \sigma_i \vec{u}_i (\vec{v}_i^\top \vec{x}) \\ &= \sum_{i=1}^{\ell} \sigma_i (\vec{v}_i \cdot \vec{x}) \vec{u}_i \text{ since } \vec{x} \cdot \vec{y} = \vec{x}^\top \vec{y} \end{aligned}$$

So, applying A to $\vec{x}$ is the same as linearly combining the $\vec{u}_i$ vectors with weights $\sigma_i(\vec{v}_i \cdot \vec{x})$. This alternative formula can provide savings when the number of nonzero σ_i values is

relatively small. More importantly, we can round small values of σ_i to zero, truncating this sum to *approximate* $A\vec{x}$ with fewer terms.

Similarly, from §6.2.1 we can write the pseudoinverse of A as:

$$A^+ = \sum_{\sigma_i \neq 0} \frac{\vec{v}_i \vec{u}_i^\top}{\sigma_i}.$$

With this formula, we can apply the same truncation trick to evaluate $A^+\vec{x}$, and we can approximate $A^+\vec{x}$ by only evaluating those terms in the sum for which σ_i is relatively *small*.

In practice, we compute the singular values σ_i as square roots of eigenvalues of $A^\top A$ or $AA^\top$, and methods like power iteration can be used to reveal a partial rather than full set of eigenvalues. Thus, if we are satisfied with approximating $A^+\vec{x}$, we can compute a few of the smallest σ_i values and truncate the formula above rather than finding A^+ completely. This also avoids ever having to compute or store the full A^+ matrix and can be accurate when A has a wide range of singular values.

Returning to our original notation $A = U\Sigma V^\top$, our argument above effectively shows that a potentially useful approximation of A is $\tilde{A} \equiv U\tilde{\Sigma}V^\top$, where $\tilde{\Sigma}$ rounds small values of Σ to zero. The column space of $\tilde{A}$ has dimension equal to the number of nonzero values on the diagonal of $\tilde{\Sigma}$. This approximation is not an *ad hoc* estimate but rather solves a difficult optimization problem posed by the following famous theorem (stated without proof):

Theorem 6.1 (Eckart-Young, 1936)**.** Suppose $\tilde{A}$ is obtained from $A = U\Sigma V^\top$ by truncating all but the k largest singular values σ_i of A to zero. Then $\tilde{A}$ minimizes both $\|A - \tilde{A}\|_{\text{Fro}}$ and $\|A - \tilde{A}\|_2$ subject to the constraint that the column space of $\tilde{A}$ has at most dimension k.

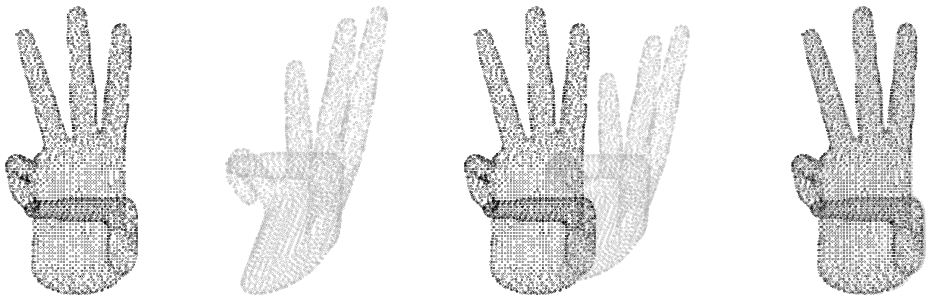
6.2.3 Matrix Norms

Constructing the SVD also enables us to return to our discussion of matrix norms from §3.3.1. For example, recall that the *Frobenius* norm of A is

$$\|A\|_{\text{Fro}}^2 \equiv \sum_{ij} a_{ij}^2.$$

If we write $A = U\Sigma V^\top$, we can simplify this expression:

$$\begin{aligned}
\|A\|_{\text{Fro}}^2 &= \sum_j \|A\vec{e}_j\|_2^2 \text{ since the product } A\vec{e}_j \text{ is the } j\text{-th column of } A \\
&= \sum_j \|U\Sigma V^\top \vec{e}_j\|_2^2, \text{ substituting the SVD} \\
&= \sum_j \vec{e}_j^\top V\Sigma^2 V^\top \vec{e}_j \text{ since } \|\vec{x}\|_2^2 = \vec{x}^\top\vec{x} \text{ and } U \text{ is orthogonal} \\
&= \|\Sigma V^\top\|_{\text{Fro}}^2 \text{ by reversing the steps above} \\
&= \|V\Sigma\|_{\text{Fro}}^2 \text{ since a matrix and its transpose have the same Frobenius norm} \\
&= \sum_j \|V\Sigma\vec{e}_j\|_2^2 = \sum_j \sigma_j^2 \|V\vec{e}_j\|_2^2 \text{ since } \Sigma \text{ is a diagonal matrix} \\
&= \sum_j \sigma_j^2 \text{ since } V \text{ is orthogonal}
\end{aligned}$$



Point cloud 1 Point cloud 2 Initial alignment Final alignment

FIGURE 6.2 If we scan a three-dimensional object from two angles, the end result is two point clouds that are not aligned. The approach explained in §6.2.4 aligns the two clouds, serving as the first step in combining the scans. Figure generated by S. Chung

Thus, the Frobenius norm of $A \in \mathbb{R}^{m \times n}$ is the sum of the squares of its singular values.

This result is of theoretical interest, but it is easier to evaluate the Frobenius norm of A by summing the squares of its elements rather than finding its SVD. More interestingly, recall that the induced two-norm of A is given by

$$\|A\|_2^2 = \max\{\lambda : \text{there exists } \vec{x} \in \mathbb{R}^n \text{ with } A^\top A\vec{x} = \lambda\vec{x}\}.$$

Now that we have studied eigenvalue problems, we realize that this value is the square root of the largest eigenvalue of $A^\top A$, or equivalently

$$\|A\|_2 = \max\{\sigma_i\}.$$

In other words, we can read the induced two-norm of A directly from its singular values.

Similarly, recall that the condition number of an invertible matrix A is given by $\text{cond } A = \|A\|_2\|A^{-1}\|_2$. By our derivation of A^+, the singular values of A^{-1} must be the reciprocals of the singular values of A. Combining this with our simplification of $\|A\|_2$ yields:

$$\text{cond } A = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

This expression provides a new formula for evaluating the conditioning of A.

There is one caveat that prevents this formula for the condition number from being used universally. Computing $\sigma_{\min}$ requires solving systems $A\vec{x} = \vec{b}$, a process which in itself may suffer from poor conditioning of A; hence, we cannot always trust our values for $\sigma_{\min}$. If this is an issue, conditioning can be bounded and approximated by using various approximations of the singular values of A.

6.2.4 The Procrustes Problem and Point Cloud Alignment

Many techniques in computer vision involve the alignment of three-dimensional shapes. For instance, suppose we have a laser scanner that collects two point clouds of the same rigid object from different views. A typical task is to align these two point clouds into a single coordinate frame, as illustrated in Figure 6.2.

Since the object is rigid, we expect there to be some rotation matrix R and translation $\vec{t} \in \mathbb{R}^3$ such that that rotating the first point cloud by R and then translating by $\vec{t}$ aligns the two data sets. Our job is to estimate $\vec{t}$ and R.

If the two scans overlap, the user or an automated system may mark n corresponding points that correspond between the two scans; we can store these in two matrices $X_1, X_2 \in \mathbb{R}^{3\times n}$. Then, for each column $\vec{x}_{1i}$ of X_1 and $\vec{x}_{2i}$ of X_2, we expect $R\vec{x}_{1i} + \vec{t} = \vec{x}_{2i}$. To account for error in measuring X_1 and X_2, rather than expecting exact equality we will minimize an energy function measuring how much this relationship holds true:

$$E \equiv \sum_i \|R\vec{x}_{1i} + \vec{t} - \vec{x}_{2i}\|_2^2.$$

If we fix R and only consider $\vec{t}$, minimizing E becomes a least-squares problem. Now, suppose we optimize for R with $\vec{t}$ fixed. This is the same as minimizing $\|RX_1 - X_2^t\|_{\text{Fro}}^2$, where the columns of X_2^t are those of X_2 translated by $\vec{t}$, subject to R being a 3×3 rotation matrix, that is, that $R^\top R = I_{3\times 3}$. This is known as the *orthogonal Procrustes problem.*

To solve this problem, we will introduce the *trace* of a square matrix as follows:

Definition 6.3 (Trace)**.** The *trace* of $A \in \mathbb{R}^{n\times n}$ is the sum of its diagonal elements:

$$\text{tr}(A) \equiv \sum_i a_{ii}.$$

In exercise 6.2 you will check that $\|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A)$. Thus, we can simplify E as follows:

$$\begin{aligned}
\|RX_1 - X_2^t\|_{\text{Fro}}^2 &= \text{tr}((RX_1 - X_2^t)^\top (RX_1 - X_2^t)) \\
&= \text{tr}(X_1^\top X_1 - X_1^\top R^\top X_2^t - X_2^{t\top} RX_1 + X_2^{t\top} X_2) \\
&= \text{const.} - 2\text{tr}(X_2^{t\top} RX_1) \\
&\qquad \text{since } \text{tr}(A+B) = \text{tr}\, A + \text{tr}\, B \text{ and } \text{tr}(A^\top) = \text{tr}(A)
\end{aligned}$$

Thus, we wish to maximize $\text{tr}(X_2^{t\top} RX_1)$ with $R^\top R = I_{3\times 3}$. From exercise 6.2, $\text{tr}(AB) = \text{tr}(BA)$. Thus our objective can simplify slightly to $\text{tr}(RC)$ with $C \equiv X_1 X_2^{t\top}$. Applying the SVD, if we decompose $C = U\Sigma V^\top$ then:

$$\begin{aligned}
\text{tr}(RC) &= \text{tr}(RU\Sigma V^\top) \text{ by definition} \\
&= \text{tr}((V^\top RU)\Sigma) \text{ since } \text{tr}(AB) = \text{tr}(BA) \\
&= \text{tr}(\tilde{R}\Sigma) \text{ if we define } \tilde{R} = V^\top RU, \text{ which is also orthogonal} \\
&= \sum_i \sigma_i \tilde{r}_{ii} \text{ since } \Sigma \text{ is diagonal}
\end{aligned}$$

Since $\tilde{R}$ is orthogonal, its columns all have unit length. This implies that $|\tilde{r}_{ii}| \leq 1$, since otherwise the norm of column i would be too big. Since $\sigma_i \geq 0$ for all i, this argument shows that we can maximize $\text{tr}(RC)$ by taking $\tilde{R} = I_{3\times 3}$, which achieves that upper bound. Undoing our substitutions shows $R = V\tilde{R}U^\top = VU^\top$.

Changing notation slightly, we have shown the following:

Theorem 6.2 (Orthogonal Procrustes)**.** The orthogonal matrix R minimizing $\|RX - Y\|_{\text{Fro}}^2$ is given by $VU^\top$, where SVD is applied to factor $XY^\top = U\Sigma V^\top$.

Returning to the alignment problem, one typical strategy employs *alternation*:

1. Fix R and minimize E with respect to $\vec{t}$.
2. Fix the resulting $\vec{t}$ and minimize E with respect to R subject to $R^\top R = I_{3\times 3}$.

3. Return to step 1.

The energy E decreases with each step and thus converges to a local minimum. Since we never optimize $\vec{t}$ and R simultaneously, we cannot guarantee that the result is the smallest possible value of E, but in practice this method works well.

6.2.5 Principal Component Analysis (PCA)

Recall the setup from §5.1.1: We wish to find a low-dimensional approximation of a set of data points stored in the columns of a matrix $X \in \mathbb{R}^{n\times k}$, for k observations in n dimensions. Previously, we showed that if we wish to project onto a single dimension, the best possible axis is given by the dominant eigenvector of $XX^\top$. With the SVD in hand, we can consider more complicated datasets that need more than one projection axis.

Suppose that we wish to choose d vectors whose span best contains the data points in X (we considered $d = 1$ in §5.1.1); we will assume $d \leq \min\{k, n\}$. We can write them in the columns of an $n \times d$ matrix C. Without affecting the column space of C, we can orthogonalize the columns of C without affecting their span. Rather than doing this operation *a posteriori*, we can safely assume ahead of time that the columns of C are orthonormal, or $C^\top C = I_{d\times d}$. Then, the projection of X onto the column space of C is given by $CC^\top X$.

Paralleling our earlier development, we can attempt to minimize $\|X - CC^\top X\|_{\text{Fro}}$ subject to $C^\top C = I_{d\times d}$. We can simplify our problem somewhat:

$$\begin{aligned}
\|X - CC^\top X\|_{\text{Fro}}^2 &= \text{tr}((X - CC^\top X)^\top (X - CC^\top X)) \text{ since } \|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A) \\
&= \text{tr}(X^\top X - 2X^\top CC^\top X + X^\top CC^\top CC^\top X) \\
&= \text{const.} - \text{tr}(X^\top CC^\top X) \text{ since } C^\top C = I_{d\times d} \\
&= -\|C^\top X\|_{\text{Fro}}^2 + \text{const.}
\end{aligned}$$

So, equivalently we can maximize $\|C^\top X\|_{\text{Fro}}^2$; for statisticians, this shows when the rows of X have mean zero that we wish to maximize the variance of the projection $C^\top X$.

Now, suppose we factor $X = U\Sigma V^\top$. Then, we wish to maximize $\|C^\top U\Sigma V^\top\|_{\text{Fro}} = \|\tilde{C}^\top \Sigma\|_{\text{Fro}} = \|\tilde{\Sigma}^\top C\|_{\text{Fro}}$ by orthogonality of V if we take $\tilde{C} = CU^\top$. If the elements of $\tilde{C}$ are $\tilde{c}_{ij}$, then expanding this norm yields

$$\|\Sigma^\top \tilde{C}\|_{\text{Fro}}^2 = \sum_i \sigma_i^2 \sum_j \tilde{c}_{ij}^2.$$

By orthogonality of the columns of $\tilde{C}$, we know $\sum_i \tilde{c}_{ij}^2 = 1$ for all j and, since $\tilde{C}$ may have fewer than n columns, $\sum_j \tilde{c}_{ij}^2 \leq 1$. Thus, the coefficient next to σ_i^2 is at most 1 in the sum above, so if we sort such that $\sigma_1 \geq \sigma_2 \geq \cdots$, then the maximum is achieved by taking the columns of $\tilde{C}$ to be $\vec{e}_1, \ldots, \vec{e}_d$. Undoing our change of coordinates, we see that our choice of C should be the first d columns of U.

We have shown that the SVD of X can be used to solve such a *principal component analysis* (PCA) problem. In practice the rows of X usually are shifted to have mean zero before carrying out the SVD.

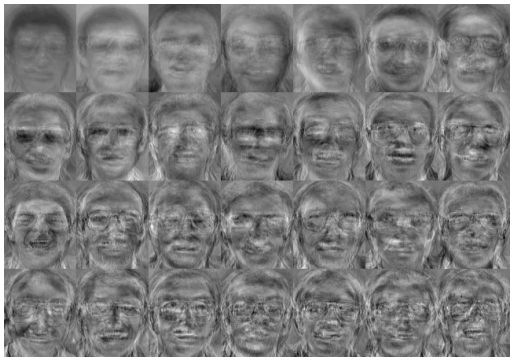
6.2.6 Eigenfaces*

One application of PCA in computer vision is the eigenfaces technique for face recognition, originally introduced in [116]. This popular method works by applying PCA to the images

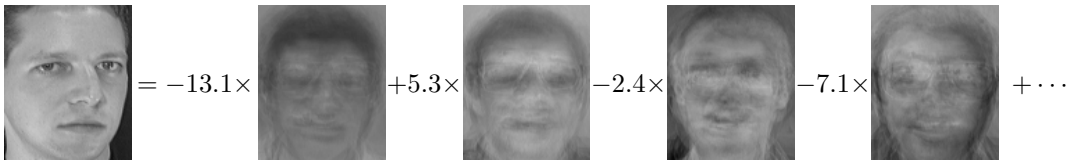
*Written with assistance by D. Hyde



(a) Input faces



(b) Eigenfaces



(c) Projection

FIGURE 6.3 The "eigenface" technique [116] performs PCA on a database of face images (a) to extract their most common modes of variation (b). For clustering, recognition, and other tasks, face images are written as linear combinations of the eigenfaces (c), and the resulting coefficients are compared. Figure generated by D. Hyde; images from the AT&T Database of Faces, AT&T Laboratories Cambridge.

in a database of faces. Projecting new input faces onto the small PCA basis encodes a face image using just a few basis coefficients without sacrificing too much accuracy, a benefit that the method inherits from PCA.

For simplicity, suppose we have a set of k photographs of faces with similar lighting and alignment, as in Figure 6.3(a). After resizing, we can assume the photos are all of size $m \times n$, so we can represent them as vectors in $\mathbb{R}^{mn}$ containing one pixel intensity per dimension. As in §6.2.5, we will store our entire database of faces in a "training matrix" $X \in \mathbb{R}^{mn \times k}$. By convention, we subtract the average face image from each column so that $X\vec{1} = \vec{0}$.

We can apply PCA to X as explained in the previous section, yielding a set of "eigenface" images in the basis matrix C representing the common modes of variation between faces. One set of eigenfaces ordered by decreasing singular value is shown in Figure 6.3(b); the first few eigenfaces alone capture common changes face shape, prominent features, and so on. Intuitively, PCA in this context searches for the most common distinguishing features that make a given face different from average.

Now that we have the eigenface basis $C \in \mathbb{R}^{mn \times d}$, we can use it for the face recognition problem. Suppose we take a new photo $\vec{x} \in \mathbb{R}^{mn}$ and wish to find the closest match in the database of faces. The projection of $\vec{x}$ onto the eigenface basis is $\vec{y} \equiv C^\top \vec{x}$. The best matching face is then the closest column of $C^\top X$, which contains the face database projected onto the eigenface basis, to $\vec{y}$.

There are two primary advantages of eigenfaces for practical face recognition. First, we usually choose $d \ll mn$, reducing the dimensionality of the search problem. More impor-

tantly, however, PCA helps separate the *relevant* modes of variation between faces from noise. In particular, differencing the mn pixels of face images independently does not search for important facial features, while the PCA axes in C are tuned to the differences observed in the columns of X.

Many modifications, improvements, and extensions have been proposed to augment the original eigenfaces technique. For example, we can set a minimum threshold so that if the input weights do not closely match any of the database weights, we report that no match was found. We also can attempt to modify PCA itself to be more sensitive to differences between identity rather than between lighting or pose. Even so, our rudimentary implementation is surprisingly effective. In our example, we train eigenfaces using photos of 40 subjects and then test using 40 *different* photos of the same subjects; the basic method described above achieves 80% recognition accuracy on this test.

6.3 EXERCISES

6.1 Suppose $A \in \mathbb{R}^{n\times n}$. Show that condition number of $A^\top A$ with respect to $\|\cdot\|_2$ is the square of the condition number of A

6.2 Suppose $A, B \in \mathbb{R}^{n\times n}$. Show $\|A\|_{\text{Fro}}^2 = \text{tr}(A^\top A)$ and $\text{tr}(AB) = \text{tr}(BA)$.

6.3 Provide the SVD and condition number with respect to $\|\cdot\|_2$ of the following matrices.

(a) $\begin{pmatrix} 0 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ \sqrt{3} & 0 & 0 \end{pmatrix}$

(b) $\begin{pmatrix} -5 \\ 3 \end{pmatrix}$

6.4 Show that $\|A\|_2 = \|\Sigma\|_2$, where $A = U\Sigma V^T$ is the singular value decomposition of A.

6.5 Show that adding a row to a matrix cannot decrease its largest or smallest singular value.

6.6 Show that the null space of a matrix $A \in \mathbb{R}^{n\times n}$ is spanned by columns of V corresponding to zero singular values, where $A = U\Sigma V^\top$ is the singular value decomposition of A.

6.7 Take $\sigma_i(A)$ to be the i-th singular value of the square matrix $A \in \mathbb{R}^{n\times n}$. Define the *nuclear norm* of A to be

$$\|A\|_* \equiv \sum_{i=1}^{n} \sigma_i(A).$$

Note: What follows is a tricky problem. Apply the mantra from this chapter: "If a linear algebra problem is hard, substitute the SVD."

(a) Show $\|A\|_* = \text{tr}(\sqrt{A^\top A})$, where trace of a matrix $\text{tr}(A)$ is the sum $\sum_i a_{ii}$ of its diagonal elements. For this problem, we will define the square root of a symmetric, positive semidefinite matrix M to be $\sqrt{M} \equiv XD^{1/2}X^\top$, where $D^{1/2}$ is the diagonal matrix containing (nonnegative) square roots of the eigenvalues of M and X contains the eigenvectors of $M = XDX^\top$.

Hint (to get started): Write $A = U\Sigma V^\top$ and argue $\Sigma^\top = \Sigma$ in this case.

(b) If $A, B \in \mathbb{R}^{n\times n}$, show $\text{tr}(AB) = \text{tr}(BA)$.

(c) Show $\|A\|_* = \max_{C^\top C=I} \text{tr}(AC)$.
Hint: Substitute the SVD of A and apply part 6.7b.

(d) Show that $\|A + B\|_* \leq \|A\|_* + \|B\|_*$.
Hint: Use part 6.7c.

(e) Minimizing $\|A\vec{x} - \vec{b}\|_2^2 + \|\vec{x}\|_1$ provides an alternative to Tikhonov regularization that can yield *sparse* vectors $\vec{x}$ under certain conditions. Assuming this is the case, explain informally why minimizing $\|A - A_0\|_{\text{Fro}}^2 + \|A\|_*$ over A for a fixed $A_0 \in \mathbb{R}^{n\times n}$ might yield a *low-rank* approximation of A_0.

(f) Provide an application of solutions to the "low-rank matrix completion" problem; 6.7e provides an optimization approach to this problem.

6.8 ("Polar decomposition") In this problem we will add one more matrix factorization to our linear algebra toolbox and derive an algorithm by N. Higham for its computation [57]. The decomposition has been used in animation applications interpolating between motions of a rigid object while projecting out undesirable shearing artifacts [106].

(a) Show that any matrix $A \in \mathbb{R}^{n\times n}$ can be factored $A = WP$, where W is orthogonal and P is symmetric and positive semidefinite. This factorization is known as the polar decomposition.
Hint: Write $A = U\Sigma V^\top$ and show $V\Sigma V^\top$ is positive semidefinite.

(b) The polar decomposition of an invertible $A \in \mathbb{R}^{n\times n}$ can be computed using an iterative scheme:

$$X_0 \equiv A \qquad\qquad X_{k+1} = \frac{1}{2}(X_k + (X_k^{-1})^\top)$$

We will prove this in a few steps:

(i) Use the SVD to write $A = U\Sigma V^\top$, and define $D_k = U^\top X_k V$. Show $D_0 = \Sigma$ and $D_{k+1} = \frac{1}{2}(D_k + (D_k^{-1})^\top)$.

(ii) From (i), each D_k is diagonal. If d_{ki} is the i-th diagonal element of D_k, show

$$d_{(k+1)i} = \frac{1}{2}\left(d_{ki} + \frac{1}{d_{ki}}\right).$$

(iii) Assume $d_{ki} \to c_i$ as $k \to \infty$ (this convergence assumption requires proof!). Show $c_i = 1$.

(iv) Use 6.8(b)iii to show $X_k \to UV^\top$.

6.9 ("Derivative of SVD," [90]) In this problem, we will continue to use the notation of problem 3.1. Our goal is to differentiate the SVD of a matrix A with respect to changes in A. Such derivatives are used to simulate the dynamics of elastic objects; see [6] for one application.

(a) Suppose $Q(t)$ is an orthogonal matrix for all $t \in \mathbb{R}$. If we define $\Omega_Q \equiv Q^\top \partial Q$, show that Ω_Q is *antisymmetric*, that is $\Omega_Q^\top = -\Omega_Q$. What are the diagonal elements of Ω_Q?

(b) Suppose for a matrix-valued function $A(t)$ we use SVD to decompose $A(t) = U(t)\Sigma(t)V(t)^\top$. Derive the following formula:

$$U^\top(\partial A)V = \Omega_U\Sigma + \partial\Sigma - \Sigma\Omega_V.$$

(c) Show how to compute $\partial\Sigma$ directly from ∂A and the SVD of A.

(d) Provide a method for finding Ω_U and Ω_V from ∂A and the SVD of A using a sequence of 2×2 solves. Conclude with formulas for ∂U and ∂V in terms of the Ω's.
Hint: It is sufficient to compute the elements of Ω_U and Ω_V above the diagonal.

6.10 ("Latent semantic analysis," [32]) In this problem, we explore the basics of *latent semantic analysis*, used in natural language processing to analyze collections of documents.

(a) Suppose we have a dictionary of m words and a collection of n documents. We can write an *occurrence matrix* $X \in \mathbb{R}^{m\times n}$ whose entries x_{ij} are equal to the number of times word i appears in document j. Propose interpretations of the entries of $XX^\top$ and $X^\top X$.

(b) Each document in X is represented using a point in $\mathbb{R}^m$, where m is potentially large. Suppose for efficiency and robustness to noise, we would prefer to use representations in $\mathbb{R}^k$, for some $k \ll m$. Apply Theorem 6.1 to propose a set of k vectors in $\mathbb{R}^m$ that best approximates the full space of documents with respect to the Frobenius norm.

(c) In *cross-language* applications, we might have a collection of n documents translated into two different languages, with m_1 and m_2 words respectively. Then, we can write two occurrence matrices $X_1 \in \mathbb{R}^{m_1\times n}$ and $X_2 \in \mathbb{R}^{m_2\times n}$. Since we do not know which words in the first language correspond to which words in the second, the columns of these these matrices are in correspondence but the rows are not.
One way to find similar phrases in the two languages is to find vectors $\vec{v}_1 \in \mathbb{R}^{m_1}$ and $\vec{v}_2 \in \mathbb{R}^{m_2}$ such that $X_1^\top\vec{v}_1$ and $X_2^\top\vec{v}_2$ are similar. To do so, we can solve a *canonical correlation* problem:

$$\max_{\vec{v}_1,\vec{v}_2} \frac{(X_1^\top\vec{v}_1)\cdot(X_2^\top\vec{v}_2)}{\|\vec{v}_1\|_2\|\vec{v}_2\|_2}.$$

Show how this minimization can be solved using eigenvector machinery.

6.11 ("Stable rank," [115]) The *stable rank* of $A \in \mathbb{R}^{n\times n}$ is defined as

$$\textsc{stable-rank}(A) \equiv \frac{\|A\|_{\text{Fro}}^2}{\|A\|_2^2}.$$

It is used in research on low-rank matrix factorization as a proxy for the rank (dimension of the column space) of A.

(a) Show that if all n columns of A are the same vector $\vec{v} \in \mathbb{R}^n\backslash\{\vec{0}\}$, then $\textsc{stable-rank}(A) = 1$.

(b) Show that when the columns of A are orthonormal, $\text{STABLE-RANK}(A) = n$.

(c) More generally, show $1 \leq \text{STABLE-RANK}(A) \leq n$.

(d) Show $\text{STABLE-RANK}(A) \leq \text{RANK}(A)$.

III

Nonlinear Techniques

CHAPTER 7

Nonlinear Systems

CONTENTS

7.1 Root-finding in a Single Variable 141
7.1.1 Characterizing Problems 141
7.1.2 Continuity and Bisection 142
7.1.3 Fixed Point Iteration 143
7.1.4 Newton's Method 145
7.1.5 Secant Method 146
7.1.6 Hybrid Techniques 147
7.1.7 Single-Variable Case: Summary 148
7.2 Multivariable Problems 148
7.2.1 Newton's Method 149
7.2.2 Making Newton Faster: Quasi-Newton and Broyden 149
7.3 Conditioning 152

TRY as we might, it is not possible to express all systems of equations in the linear framework we have developed over the last several chapters. Logarithms, exponentials, trigonometric functions, absolute values, polynomials, and so on are commonplace in practical problems, but except in a few special cases none of these functions is linear. When these functions appear, we must employ a more general—but often less efficient—toolbox for nonlinear problems.

7.1 ROOT-FINDING IN A SINGLE VARIABLE

We begin our discussion by considering methods for root-finding in a single scalar variable. Given a function $f(x) : \mathbb{R} \to \mathbb{R}$, we wish to develop algorithms for finding points $x^* \in \mathbb{R}$ such that $f(x^*) = 0$; we call x^* a *root* of f. Single-variable problems in linear algebra are not particularly interesting; after all we can solve the equation $ax - b = 0$ in closed form as $x^* = b/a$. Roots of a nonlinear equation like $y^2 + e^{\cos y} - 3 = 0$, however, are less easily calculated.

7.1.1 Characterizing Problems

We no longer can assume f is linear, but without *any* information about its structure we are unlikely to make headway on finding a root of f. For instance, a solver is guaranteed to fail finding zeros of $f(x)$ given by

$$f(x) = \begin{cases} -1 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Or even more deviously:

$$f(x) = \begin{cases} -1 & x \in \mathbb{Q} \\ 1 & \text{otherwise} \end{cases}$$

These examples are trivial in the sense that any reasonable client of root-finding software would be unlikely to expect it to succeed in this case, but more subtle examples are not much more difficult to construct.

For this reason, we must add some "regularizing" assumptions about f providing a toehold into the possibility of designing root-finding techniques. Typical such assumptions are below:

- *Continuity:* A function f is *continuous* if it can be drawn without lifting up a pen; more formally, f is continuous if the difference $f(x) - f(y)$ vanishes as $x \to y$.
- *Lipschitz:* A function f is *Lipschitz continuous* if there exists a constant c such that $|f(x) - f(y)| \leq c|x - y|$; Lipschitz functions need not be differentiable but are limited in their rates of change.
- *Differentiability:* A function f is *differentiable* if its derivative f' exists for all x.
- C^k*:* A function is C^k if it is differentiable k times and each of those k derivatives is continuous; C^∞ indicates that all derivatives of f exist and are continuous.

Example 7.1 (Classifying functions)**.** The function $f(x) = \cos x$ is C^∞ and Lipschitz on $\mathbb{R}$. The function $g(x) = x^2$ as a function on $\mathbb{R}$ is C^∞ but *not* Lipschitz. In particular $|f(x) - f(0)| = x^2$, which cannot be bounded by any linear function of x as $x \to \infty$. When restricted to the unit interval $[0, 1]$, however, $g(x)$ can be considered Lipschitz since its slope is bounded by 2 in this interval; we say f is "locally Lipschitz" since this property holds on any interval $[a, b]$. The function $h(x) = |x|$ is continuous—or C^0—and Lipschitz but not differentiable thanks to its singularity at $x = 0$.

As we add stronger and stronger assumptions about f, we can design more effective algorithms to solve $f(x^*) = 0$. We will illustrate this effect by considering a few algorithms below.

7.1.2 Continuity and Bisection

Suppose all we know about f is that it is continuous. In this case, we can state an intuitive theorem from standard single-variable calculus:

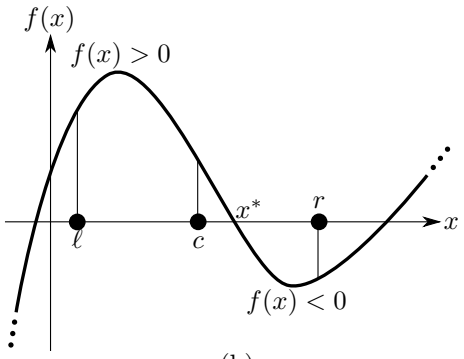
Theorem 7.1 (Intermediate Value Theorem)**.** Suppose $f : [a, b] \to \mathbb{R}$ is continuous. Suppose $f(x) < u < f(y)$. Then, there exists z between x and y such that $f(z) = u$.

In other words, the function f must achieve every value in between $f(x)$ and $f(y)$.

Suppose we are given as input the function f as well as two values ℓ and r such that $f(\ell) \cdot f(r) < 0$, that is, $f(\ell)$ and $f(r)$ have opposite sign. Then, by the Intermediate Value Theorem we know that somewhere between ℓ and r there is a root of f. Similar to binary search, this property provides a bisection strategy for finding x^*, shown in Figure 7.1. This algorithm divides the interval $[\ell, r]$ in half recursively, each time keeping the side in which a root is known to exist by the Intermediate Value Theorem. It converges *unconditionally*, in the sense that so long as $f(\ell) \cdot f(r) < 0$ eventually both ℓ and r are guaranteed to become arbitrarily close to one another and converge to a root x^* of $f(x)$.

```
function BISECTION(f(x), ℓ, r)
  for k ← 1, 2, 3, ...
    c ← (ℓ+r)/2
    if |f(c)| < ε_f or |r − ℓ| < ε_x then
      return x* ≈ c
    else if f(ℓ) · f(c) < 0 then
      r ← c
    else
      ℓ ← c
```

(a)



(b)

FIGURE 7.1 Pseudocode (a) and an illustration of (b) the bisection algorithm for finding roots of continuous $f(x)$ given endpoints $\ell, r \in \mathbb{R}$ with $f(\ell) \cdot f(r) < 0$. The interval $[c, r]$ contains a root x^* because $f(c)$ and $f(r)$ have opposite sign.

Analysis of Bisection. Bisection is the simplest but not necessarily the fastest technique for root-finding. As with eigenvalue methods, bisection inherently is iterative and may never provide an *exact* solution x^*; this property is true for nearly any root-finding algorithm unless we put strong assumptions on f. We can ask, however, how close the value c_k of c in the k-th iteration is to the root x^* that we hope to compute. This analysis will provide a baseline for comparison to other methods.

In general, suppose we can establish an error bound E_k such that the estimate x_k of the root x^* during the k-th iteration of a root-finding method satisfies $|x_k - x^*| < E_k$. Any algorithm with $E_k \to 0$ represents a *convergent* scheme; the speed of convergence, however, can be characterized by the rate at which E_k approaches 0.

For example, in bisection since both c_k and x^* are in the interval $[\ell_k, r_k]$, an upper bound of error is given by $E_k \equiv |r_k - \ell_k|$. Since we divide the interval in half each iteration, we can reduce our error bound by half in each iteration: $E_{k+1} = {}^1/_2 E_k$. Since E_{k+1} is linear in E_k, we say that bisection exhibits *linear* convergence.

In exchange for unconditional linear convergence, bisection requires initial estimates of ℓ and r bracketing a root. While some heuristic search methods exist for finding these points, unless more is known about the form of f, finding this pair may be nearly as difficult as computing a root! In this case, bisection might be thought of as a method for *refining* a root estimate rather than for global search.

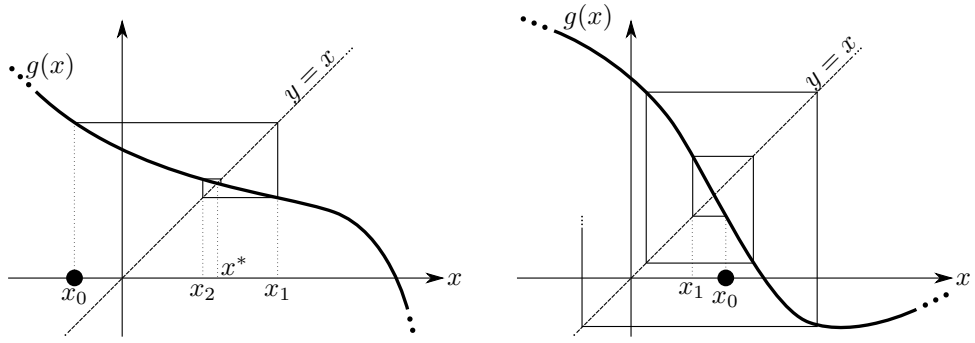
7.1.3 Fixed Point Iteration

Bisection is guaranteed to converge to a root of any continuous function f, but if we know more about f we can formulate algorithms that can converge more quickly.

As an example, suppose we wish to find x^* satisfying $g(x^*) = x^*$; this setup is equivalent to root-finding since solving $g(x^*) = x^*$ is the same as solving $g(x^*) - x^* = 0$. As an additional piece of information, however, we also might know that g is Lipschitz with constant $0 \le c < 1$ (see §7.1.1). This condition defines g as a *contraction*, since $|g(x) - g(y)| < |x - y|$ for any x, y.

The system $g(x) = x$ suggests a potential solution method:

1. Take x_0 to be an initial guess of a root.



(a) Convergence (b) Divergence

FIGURE 7.2 Convergence of fixed point iteration. Fixed-point iteration searches for the intersection of $g(x)$ with the line $y = x$ by iterating $x_k = g(x_{k-1})$. One way to visualize this method on the graph of $g(x)$ visualized above is that it alternates between moving horizontally to the line $y = x$ and vertically to the position $g(x)$. Fixed point iteration converges (a) when the slope of $g(x)$ is small and diverges (b) otherwise.

2. Iterate $x_k = g(x_{k-1})$.

If this iteration converges, the result is a *fixed point* of g satisfying the criteria above.

When $c < 1$, the Lipschitz property ensures convergence to a root if one exists. If we take $E_k = |x_k - x^*|$, then we have the following property:

$$\begin{aligned} E_k &= |x_k - x^*| \\ &= |g(x_{k-1}) - g(x^*)| \text{ by design of the iterative scheme and definition of } x^* \\ &\leq c|x_{k-1} - x^*| \text{ since } g \text{ is Lipschitz} \\ &= cE_{k-1} \end{aligned}$$

Applying this statement inductively shows $E_k \leq c^k|E_0| \to 0$ as $k \to \infty$. Thus, for this choice of c, fixed point iteration converges to the desired x^*.

If g is Lipschitz with constant $c < 1$ in a *neighborhood* $[x^* - \delta, x^* + \delta]$, then so long as x_0 is chosen in this interval, fixed point iteration will converge. This is true since our expression for E_k above shows that it shrinks each iteration. When the Lipschitz constant is too large—or equivalently, when g has large slope—fixed point iteration diverges. Figure 7.2 visualizes the two possibilities.

One important case occurs when g is C^1 and $|g'(x^*)| < 1$. By continuity of g' in this case, there are values $\varepsilon, \delta > 0$ such that $|g'(x)| < 1 - \varepsilon$ for any $x \in (x^* - \delta, x^* + \delta)$.* Take any $x, y \in (x^* - \delta, x^* + \delta)$. Then, we have

$$\begin{aligned} |g(x) - g(y)| &= |g'(\theta)| \cdot |x - y| \text{ by the Mean Value Theorem, for some } \theta \in [x, y] \\ &< (1 - \varepsilon)|x - y| \end{aligned}$$

This argument shows that g is Lipschitz with constant $1-\varepsilon < 1$ in the interval $(x^*-\delta, x^*+\delta)$. Thus, when g is continuously differentiable and $g'(x^*) < 1$, fixed point iteration will converge to x^* when the initial guess x_0 is close by.

*This statement is hard to parse: Make sure you understand it!

So far we have little reason to use fixed point iteration: We have shown it is guaranteed to converge only when g is Lipschitz, and our argument about the E_k's shows linear convergence like bisection. There is one case, however, in which fixed point iteration provides an advantage.

Suppose g is differentiable with $g'(x^*) = 0$. Then, the first-order term vanishes in the Taylor series for g, leaving behind:

$$g(x_k) = g(x^*) + \frac{1}{2}g''(x^*)(x_k - x^*)^2 + O\left((x_k - x^*)^3\right).$$

In this case we have:

$$\begin{aligned}
E_k &= |x_k - x^*| \\
&= |g(x_{k-1}) - g(x^*)| \text{ as before} \\
&= \frac{1}{2}|g''(x^*)|(x_{k-1} - x^*)^2 + O((x_{k-1} - x^*)^3) \text{ from the Taylor argument} \\
&\leq \frac{1}{2}(|g''(x^*)| + \varepsilon)|(x_{k-1} - x^*)^2 \text{ for some } \varepsilon \text{ so long as } x_{k-1} \text{ is close to } x^* \\
&= \frac{1}{2}(|g''(x^*)| + \varepsilon)E_{k-1}^2
\end{aligned}$$

By this chain of inequalities, in this case E_k is *quadratic* in E_{k-1}, so we say fixed point iteration can have *quadratic convergence*. This implies that $E_k \to 0$ much faster, so we will need fewer iterations to reach a reasonable root.

7.1.4 Newton's Method

Our discussion of fixed point iteration showed a few cases in which knowing a function is differentiable helps us find its roots more quickly. With this observation in mind, we tighten our class of functions once more to derive a method based more fundamentally on the differentiability assumption that has consistent quadratic convergence. In particular, we will again suppose that we wish to solve $f(x^*) = 0$, but now we assume that $f \in C^1$, a slightly tighter condition than Lipschitz.

At $x_k \in \mathbb{R}$, since f is differentiable we can approximate it using a tangent line:

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k)$$

Solving this approximation for $f(x) \approx 0$ yields an approximation of the root x_{k+1}:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

In reality, x_{k+1} may not satisfy $f(x_{k+1}) = 0$, but since it is the root of an approximation of f we might hope that it is closer to x^* than x_k. If this is true, than iterating this formula should give x_k's that get closer and closer to x^*. This iterative technique is known as *Newton's method* for root-finding, and it amounts to repeatedly solving linear approximations of the original nonlinear problem. It is illustrated in Figure 7.3.

If we define

$$g(x) = x - \frac{f(x)}{f'(x)},$$

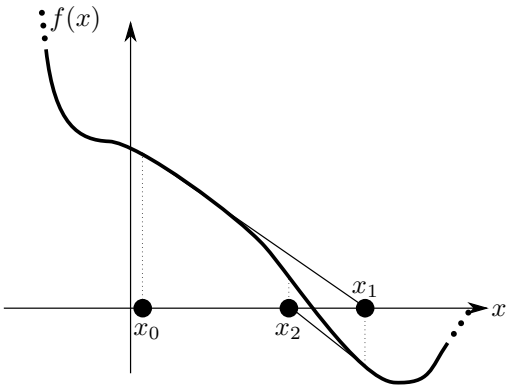


FIGURE 7.3 Newton's method iteratively approximates $f(x)$ with tangent lines to find roots of a differentiable function $f(x)$.

then Newton's method amounts to fixed point iteration on g. Differentiating,

$$\begin{aligned} g'(x) &= 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} \text{ by the quotient rule} \\ &= \frac{f(x)f''(x)}{f'(x)^2} \end{aligned}$$

Suppose x^* is a *simple* root, meaning $f'(x^*) \neq 0$. Then, $g'(x^*) = 0$, and by our analysis of fixed-point iteration in §7.1.3, Newton's method must converge *quadratically* to x^* when we start from a sufficiently close initial guess x_0. When x^* is not simple, however, convergence can be linear or worse.

The derivation of Newton's method via linear approximation suggests other methods using more terms in the Taylor series. For instance, "Halley's method" adds terms involving f'' via quadratic approximation, and more general "Householder methods" can include an arbitrary number of derivatives. These techniques offer higher-order convergence at the cost of having to evaluate high-order derivatives and the possibility of more exotic failure modes. Other methods replace Taylor series with alternative approximations; for example, "linear fractional interpolation" uses rational functions to better approximate functions with asymptotes.

7.1.5 Secant Method

One concern about Newton's method is the cost of evaluating f and its derivative f'. If f is complicated, we may wish to minimize the number of times we have to evaluate either of these functions. Higher orders of convergence for root-finding alleviate this problem by reducing the number of iterations needed to approximate x^*, but we also can design numerical methods that explicitly avoid evaluating costly derivatives.

Example 7.2 (Rocket design)**.** Suppose we are designing a rocket and wish to know how much fuel to add to the engine. For a given number of gallons x, we can write a function $f(x)$ giving the maximum height of the rocket during flight; our engineers have specified that the rocket should reach a height h, so we need to solve $f(x) = h$. Evaluating $f(x)$ involves simulating a rocket as it takes off and monitoring its fuel consumption, which

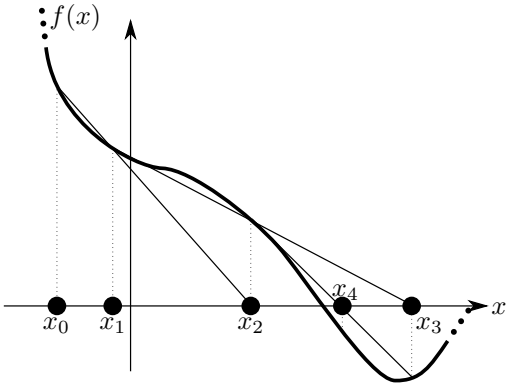


FIGURE 7.4 The secant method is similar to Newton's method (Figure 7.3) but approximates tangents to $f(x)$ as the lines through previous iterates. It requires both x_0 and x_1 for initialization.

is an expensive proposition. Even if f is differentiable, we might not be able to evaluate f' in a practical amount of time.

One strategy for designing lower-impact methods is to reuse data as much as possible. For instance, we could approximate the derivative f' appearing in Newton's method as follows:

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Since we had to compute $f(x_{k-1})$ in the previous iteration anyway, we reuse this value to approximate the derivative for the next one. This approximation works well when x_k's are near convergence and close to one another. Plugging our approximation for f' into Newton's method results in a new scheme known as the *secant method*, illustrated in Figure 7.4:

$$x_{k+1} = x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})}$$

The user will have to provide two initial guesses x_0 and x_1 to start this scheme, or can run a single iteration of Newton to get it started.

Analyzing the secant method is more involved than the other methods we have considered because it uses both $f(x_k)$ and $f(x_{k-1})$; proof of its convergence is outside the scope of our discussion. Error analysis reveals that the secant method decreases error at a rate of $1+\sqrt{5}/2$ (the "Golden Ratio"), which is between linear and quadratic. Since convergence is *close* to that of Newton's method without the need for evaluating f', the secant method is a strong alternative.

7.1.6 Hybrid Techniques

With additional engineering, we can combine the advantages of different root-finding algorithms. For instance, we might make the following observations:

- *Bisection* is guaranteed to converge, but only at a linear rate.
- The *secant method* has a faster rate of convergence, but it may not converge at all if the initial guess x_0 is far from the root x^*.

Suppose we have bracketed a root of $f(x)$ in the interval $[\ell_k, r_k]$. Given the iterates x_k and x_{k-1}, we could take the next estimate x_{k+1} to be either of the following:

- The next secant method iterate, if it is contained in (ℓ_k, r_k).
- The midpoint $\ell_k + r_k/2$ otherwise.

This combination of the secant and midpoint methods guarantees that $x_{k+1} \in (\ell_k, r_k)$. Regardless of the choice above, we can update the bracket containing the root to $[\ell_{k+1}, r_{k+1}]$ as in bisection by examining the sign of $f(x_{k+1})$.

The algorithm above, called "Dekker's method," attempts to combine the unconditional convergence of bisection with the stronger root estimates of the secant method. In many cases it is successful, but its convergence rate is somewhat difficult to analyze. Specialized failure modes can reduce this method to linear convergence or worse: In some cases, bisection can converge *more* quickly! Other techniques, e.g. "Brent's method," make bisection steps more often to strengthen convergence and can exhibit guaranteed behavior at the cost of a more complex implementation.

7.1.7 Single-Variable Case: Summary

We have considered a number of methods for solving $f(x^*) = 0$ in the single-variable case. We only have scratched the surface of the set of such techniques; many other iterative schemes for root-finding exist, with different guarantees, convergence rates, and caveats. Regardless, we can make a number of observations:

- To support arbitrary functions f that may not have closed-form solutions to $f(x^*) = 0$, we adopt iterative schemes that get closer and closer to the desired root.
- We wish for the sequence x_k of root estimates to reach x^* as quickly as possible. If E_k is an error bound, then we can characterize a number of convergence situations assuming $E_k \to 0$ as $k \to \infty$. Some conditions that may hold for large k are:
  1. Linear convergence: $E_{k+1} \leq CE_k$ for some $C < 1$
  2. Superlinear convergence: $E_{k+1} \leq CE_k^r$ for $r > 1$ (now we do not require $C < 1$ since if E_k is small enough, the r power can cancel the effects of C)
  3. Quadratic convergence: $E_{k+1} \leq CE_k^2$
  4. Cubic convergence: $E_{k+1} \leq CE_k^3$ (and so on)
- A method might converge more quickly, needing fewer iterations to get sufficiently close to x^*, but during each individual iteration it may require additional computation time. For this reason, it may be preferable to do more iterations of a simpler method than fewer iterations of a more complex one.

7.2 MULTIVARIABLE PROBLEMS

Some applications may require solving the multivariable problem $f(\vec{x}) = \vec{0}$ for a function $f : \mathbb{R}^n \to \mathbb{R}^m$. We have already seen one instance of this problem when solving $A\vec{x} = \vec{b}$, which is equivalent to finding roots of $f(\vec{x}) \equiv A\vec{x} - \vec{b}$, but the general case is considerably more difficult. In particular, strategies like bisection are difficult to extend since we now must guarantee that m different values are all zero *simultaneously*.

7.2.1 Newton's Method

Thankfully, one of our single-variable strategies extends in a straightforward way. Recall from §0.4.2 that for $f : \mathbb{R}^n \to \mathbb{R}^m$ we can write the *Jacobian* matrix, which gives the derivative of each component of f in each of the coordinate directions:

$$(Df)_{ij} \equiv \frac{\partial f_i}{\partial x_j}$$

We can use the Jacobian of f to extend our derivation of Newton's method to multiple dimensions. In particular, the first-order approximation of f is given by:

$$f(\vec{x}) \approx f(\vec{x}_k) + Df(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k).$$

Substituting the desired condition $f(\vec{x}) = \vec{0}$ yields the following linear system for the next iterate $\vec{x}_{k+1}$:

$$Df(\vec{x}_k) \cdot (\vec{x}_{k+1} - \vec{x}_k) = -f(\vec{x}_k)$$

When Df is square and invertible, corresponding to $f : \mathbb{R}^n \to \mathbb{R}^n$, we obtain the typical iteration for Newton's method:

$$\vec{x}_{k+1} = \vec{x}_k - [Df(\vec{x}_k)]^{-1} f(\vec{x}_k),$$

where as always we do not explicitly compute the matrix $[Df(\vec{x}_k)]^{-1}$ but rather solve a linear system, e.g. using the techniques from Chapter 2. Otherwise, this equation can be solved using the pseudoinverse when $m < n$ to find one of likely many roots of f; when $m > n$ one can attempt least-squares, but the existence of a root and convergence of this technique are both unlikely.

An analogous multidimensional argument to that in §7.1.3 shows that fixed-point methods like Newton's method iterating $\vec{x}_{k+1} = g(\vec{x}_k)$ converge when the largest-magnitude eigenvalue of Dg is less than 1 (exercise 7.1). Then, an argument identical to the one-dimensional case in §7.1.4 shows that Newton's method in multiple variables can have quadratic convergence near roots $\vec{x}^*$ for which $Df(\vec{x}^*)$ is nonsingular.

7.2.2 Making Newton Faster: Quasi-Newton and Broyden

As m and n increase, Newton's method becomes very expensive. For each iteration, a *different* matrix $Df(\vec{x}_k)$ must be inverted; because it changes each time, factoring $Df(\vec{x}_k) = L_k U_k$ only adds cost.

Quasi-Newton algorithms apply various approximations to reduce the cost of individual iterations. One approach extends our derivation of the secant method beyond one dimension. Just as the secant method contains the same division operation as Newton's method, such secant-like approximations will not necessarily alleviate the need to invert a matrix, but they do make it possible to carry out root-finding without explicitly calculating the Jacobian Df. Such an extension to multiple dimensions is not obvious, however, since divided differences yield a single value rather than a full approximate Jacobian matrix.

The directional derivative of f in the direction $\vec{v}$ is given by $D_{\vec{v}} f = Df \cdot \vec{v}$. To imitate the secant method, we can use this scalar formula to our advantage by requiring that our approximation J of a Jacobian satisfy

$$J_k \cdot (\vec{x}_k - \vec{x}_{k-1}) = f(\vec{x}_k) - f(\vec{x}_{k-1}).$$

This formula, however, does not determine the action of J on any vector perpendicular

to $\vec{x}_k - \vec{x}_{k-1}$, so we need additional approximation assumptions to describe a complete root-finding algorithm.

One algorithm using the approximation above is *Broyden's method*, which maintains not only an estimate $\vec{x}_k$ of $\vec{x}^*$ but also a full matrix J_k estimating the Jacobian at $\vec{x}_k$ satisfying the condition above. Initial estimates J_0 and $\vec{x}_0$ both must be supplied by the user; commonly we approximate $J_0 = I_{n\times n}$.

Suppose we have an estimate J_{k-1} of the Jacobian left over from the previous iteration. We now have a new data point $\vec{x}_k$ at which we have evaluated $f(\vec{x}_k)$, so we would like to update J_{k-1} to a new matrix J_k taking into account this new piece of information. Broyden's method applies the directional derivative approximation above while requiring that the new approximate Jacobian J_k is as similar as possible to J_{k-1} by solving the following optimization problem:

$$\begin{array}{rl} \text{minimize}_{J_k} & \|J_k - J_{k-1}\|_{\text{Fro}}^2 \\ \text{such that} & J_k \cdot (\vec{x}_k - \vec{x}_{k-1}) = f(\vec{x}_k) - f(\vec{x}_{k-1}) \end{array}$$

To solve this problem, define $\Delta J \equiv J_k - J_{k-1}$, $\Delta\vec{x} \equiv \vec{x}_k - \vec{x}_{k-1}$, and $\vec{d} \equiv f(\vec{x}_k) - f(\vec{x}_{k-1}) - J_{k-1} \cdot \Delta\vec{x}$. Making these substitutions yields the following form:

$$\begin{array}{rl} \text{minimize}_{\Delta J} & \|\Delta J\|_{\text{Fro}}^2 \\ \text{such that} & \Delta J \cdot \Delta\vec{x} = \vec{d} \end{array}$$

If we take $\vec{\lambda}$ to be a Lagrange multiplier, this minimization is equivalent to finding critical points of the Lagrangian Λ:

$$\Lambda = \|\Delta J\|_{\text{Fro}}^2 + \vec{\lambda}^\top(\Delta J \cdot \Delta\vec{x} - \vec{d})$$

Differentiating with respect to an unknown element $(\Delta J)_{ij}$ shows:

$$0 = \frac{\partial\Lambda}{\partial(\Delta J)_{ij}} = 2(\Delta J)_{ij} + \lambda_i(\Delta\vec{x})_j \implies \Delta J = -\frac{1}{2}\vec{\lambda}(\Delta\vec{x})^\top$$

Substituting into $\Delta J \cdot \Delta\vec{x} = \vec{d}$ shows $\vec{\lambda}(\Delta\vec{x})^\top(\Delta\vec{x}) = -2\vec{d}$, or equivalently $\vec{\lambda} = -2\vec{d}/\|\Delta\vec{x}\|_2^2$. Finally, we can substitute into the Lagrange multiplier expression to find:

$$\Delta J = -\frac{1}{2}\vec{\lambda}(\Delta\vec{x})^\top = \frac{\vec{d}(\Delta\vec{x})^\top}{\|\Delta x\|_2^2}$$

Expanding out our substitution shows:

$$\begin{aligned} J_k &= J_{k-1} + \Delta J \\ &= J_{k-1} + \frac{\vec{d}(\Delta\vec{x})^\top}{\|\Delta x\|_2^2} \\ &= J_{k-1} + \frac{(f(\vec{x}_k) - f(\vec{x}_{k-1}) - J_{k-1}\cdot\Delta\vec{x})}{\|\vec{x}_k - \vec{x}_{k-1}\|_2^2}(\vec{x}_k - \vec{x}_{k-1})^\top \end{aligned}$$

Broyden's method alternates between this update and the corresponding Newton step $\vec{x}_{k+1} = \vec{x}_k - J_k^{-1}f(\vec{x}_k)$. Additional efficiency in some cases can be gained by keeping track of the matrix J_k^{-1} explicitly rather than the matrix J_k, which can be updated using a similar formula; this possibility is explored via the "Sherman-Morrison update formula" in exercise 7.6. Both versions of the algorithm are shown in Figure 7.5.

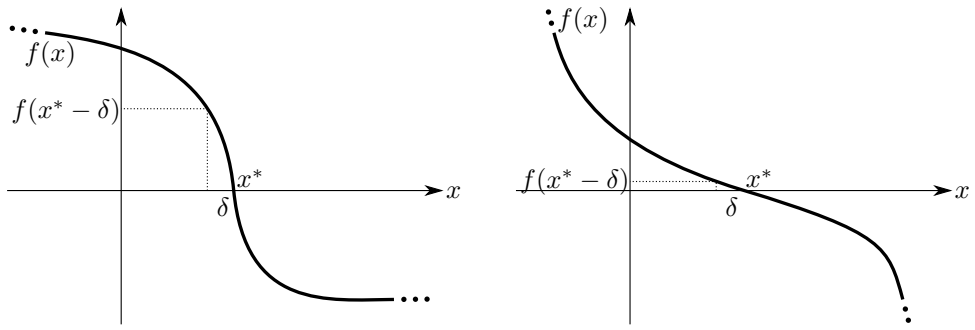
function BROYDEN($f(\vec{x}), \vec{x}_0, J_0$)
  $J \leftarrow J_0$ ▷ Can default to $I_{n\times n}$
  $\vec{x} \leftarrow \vec{x}_0$
  for $k \leftarrow 1, 2, 3, \ldots$
    $\Delta\vec{x} \leftarrow -J^{-1}f(\vec{x})$ ▷ Linear
    $\Delta f \leftarrow f(\vec{x} + \Delta x) - f(\vec{x})$
    $\vec{x} \leftarrow \vec{x} + \Delta\vec{x}$
    $J \leftarrow J + \frac{(\Delta f - J\Delta\vec{x})}{\|\Delta\vec{x}\|_2^2}(\Delta x)^\top$
  return $\vec{x}$

(a)

function BROYDEN-INVERTED($f(\vec{x}), \vec{x}_0, J_0^{-1}$)
  $J^{-1} \leftarrow J_0^{-1}$ ▷ Can default to $I_{n\times n}$
  $\vec{x} \leftarrow \vec{x}_0$
  for $k \leftarrow 1, 2, 3, \ldots$
    $\Delta\vec{x} \leftarrow -J^{-1}f(\vec{x})$ ▷ Matrix multiply
    $\Delta f \leftarrow f(\vec{x} + \Delta x) - f(\vec{x})$
    $\vec{x} \leftarrow \vec{x} + \Delta\vec{x}$
    $J^{-1} \leftarrow J^{-1} + \frac{\Delta\vec{x} - J^{-1}\Delta f}{\|\Delta f\|_2^2}(\Delta f)^\top$
  return $\vec{x}$

(b)

FIGURE 7.5 Broyden's method as described in §7.2.2 requires solving a linear system of equations (a), but after applying the formula from exercise 7.6 yields an equivalent method using only matrix multiplies by updating the inverse matrix J^{-1} directly instead of J (b).



(a) Good conditioning (b) Poor conditioning

FIGURE 7.6 Intuition for the conditioning of finding roots of a function $f(x)$. When the slope at the root x^* is large, the problem is well-conditioned because moving a small distance δ away from x^* makes the value of f change by a large amount (a). When the slope at x^* is smaller, values of $f(x)$ remain close to zero as we move away from the root, making it harder to pinpoint the exact location of x^* (b).

7.3 CONDITIONING

We already showed in Example 1.9 that the condition number of root-finding in a single variable is:

$$\text{cond}_{x^*} f = \frac{1}{|f'(x^*)|}$$

As illustrated in Figure 7.6, this condition number shows that the best possible situation for root-finding occurs when f is changing rapidly near x^*, since in this case perturbing x^* will make f take values far from 0.

Applying an identical argument when f is multidimensional gives a condition number of $\|Df(\vec{x}^*)\|^{-1}$. When Df is not invertible, the condition number is *infinite*. This degeneracy occurs because perturbing $\vec{x}^*$ preserves $f(\vec{x}) = \vec{0}$ to first order, and indeed such a condition can create challenging root-finding cases exaggerating that shown in Figure 7.6(b).

7.4 EXERCISES

7.1[DH] Recall from §7.1.3 the proof of conditions under which single-variable fixed point iteration converges. Consider now the multivariable fixed point iteration scheme $\vec{x}_{k+1} \equiv g(\vec{x}_k)$ for $g : \mathbb{R}^n \to \mathbb{R}^n$.

(a) Suppose that $g \in \mathcal{C}^1$ and that $\vec{x}_k$ is within a small neighborhood of a fixed point $\vec{x}^*$ of g. Suggest a condition on the Jacobian Dg of g that guarantees g is Lipschitz in this neighborhood.

(b) Using the previous result, derive a bound for the error of $\vec{x}_{k+1}$ in terms of the error of $\vec{x}_k$ and the Jacobian of g.

(c) Show a condition on the eigenvalues of Dg that guarantees convergence of multivariable fixed point iteration.

(d) How does the rate of convergence change if $Dg(\vec{x}^*) = 0$?

7.2[DH] Which method would you recommend for finding the root of $f : \mathbb{R} \to \mathbb{R}$ if all you know about f is that:

(a) $f \in C^1$ and f' is inexpensive to evaluate

(b) f is Lipschitz with constant c satisfying $0 \leq c \leq 1$

(c) $f \in C^1$ and f' is costly to evaluate

(d) $f \in C^0 \backslash C^1$, the continuous but non-differentiable functions

7.3[DH] Provide an example of root-finding problems that satisfy the following criteria:

(a) Can be solved by bisection but not by fixed-point iteration

(b) Can be solved using fixed-point iteration, but not using Newton's method

7.4 Is Newton's method guaranteed to have quadratic convergence? Why?

7.5[DH] Suppose we wish to compute $\sqrt[n]{y}$ for a given $y \geq 0$. Using the techniques from this chapter, derive a quadratically convergent iterative method that finds this root.

7.6 As promised, in this problem we show how to carry out Broyden's method for finding roots without solving linear systems of equations.

(a) Verify the Sherman-Morrison formula, for invertible $A \in \mathbb{R}^{n\times n}$ and vectors $\vec{u}, \vec{v} \in \mathbb{R}^n$:

$$(A + \vec{u}\vec{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\vec{u}\vec{v}^\top A^{-1}}{1 + \vec{v}^\top A^{-1}\vec{u}}$$

(b) Use this formula to show that the algorithm in Figure 7.5(b) is equivalent to Broyden's method as described in §7.2.2.

7.7 In this problem, we will derive a technique is known as Newton-Raphson division. Thanks to its fast convergence, it is often implemented in hardware for IEEE-754 floating point arithmetic.

(a) Show how the reciprocal $\frac{1}{a}$ of $a \in \mathbb{R}$ can be computed iteratively using Newton's method. Write your iterative formula in a way that requires at most two multiplications, one addition or subtraction, and no divisions.

(b) Take x_k to be the estimate of $\frac{1}{a}$ during the k-th iteration of Newton's method. If we define $\varepsilon_k \equiv ax_k - 1$, show that $\varepsilon_{k+1} = -\varepsilon_k^2$.

(c) Approximately how many iterations of Newton's method are needed to compute $\frac{1}{a}$ within d binary decimal points? Write your answer in terms of ε_0 and d, and assume $|\varepsilon_0| < 1$.

(d) Is this method always convergent regardless of the initial guess of $\frac{1}{a}$?

7.8 In this problem, we will develop a method for solving least-squares with quadratic inequality (LSQI, [46])

$$\min_{\|\vec{x}\|_2 \le 1} \|A\vec{x} - \vec{b}\|_2.$$

You can assume $A \in \mathbb{R}^{m\times n}$ is overdetermined.

(a) The optimal $\vec{x}$ either satisfies $\|\vec{x}\|_2 < 1$ or $\|\vec{x}\|_2 = 1$. Explain how to distinguish between the two cases; when $\|\vec{x}\|_2 < 1$, give a formula for $\vec{x}$.

(b) Suppose we are in the $\|\vec{x}\|_2 = 1$ case. Show that there exists $\lambda \in \mathbb{R}$ such that $(A^\top A + \lambda I_{n\times n})\vec{x} = A^\top \vec{b}$.

(c) Define $f(\lambda) \equiv \|\vec{x}(\lambda)\|_2^2 - 1$, where $\vec{x}(\lambda)$ is the solution to the system from part 7.8b. Show $f(0) > 0$ and that $f(\lambda) < 0$ for sufficiently large $\lambda > 0$.

(d) Propose a strategy for the $\|\vec{x}\|_2 = 1$ case using root-finding.

7.9 (Proposed by A. Nguyen) Suppose we have a polynomial $p(x) = a_k x^k + \cdots + a_1 x + a_0$. You can assume $p(x)$ is not the zero polynomial $q(x) \equiv 0$.

(a) Suppose the derivative $p'(x)$ has no roots in (a, b). How many roots can $p(x)$ have in this interval?

(b) Using the result of part 7.9a, propose a recursive algorithm for estimating all the real roots of $p(x)$. Assume we know that the roots of $p(x)$ are at least ε apart.

(c) Discuss the numerical and efficiency properties of your technique. What can happen if ε is unknown?

7.10 Root-finding for complex- or real-valued polynomials is closely linked to the eigenvalue problem considered in Chapter 5.

(a) Give a matrix A whose eigenvalues are the roots of a given polynomial $p(x) = a_k x^k + \cdots + a_1 x + a_0$.

(b) Show that the eigenvalues of a matrix $A \in \mathbb{R}^{n \times n}$ are the roots of a polynomial function. Is it advisable to use root-finding algorithms from this chapter for the eigenvalue problem?

CHAPTER 8

Unconstrained Optimization

CONTENTS

8.1 Unconstrained Optimization: Motivation 155
8.2 Optimality 157
8.2.1 Differential Optimality 158
8.2.2 Alternative Conditions for Optimality 160
8.3 One-Dimensional Strategies 161
8.3.1 Newton's Method 162
8.3.2 Golden Section Search 162
8.4 Multivariable Strategies 165
8.4.1 Gradient Descent 165
8.4.2 Newton's Method in Multiple Variables 166
8.4.3 Optimization without Hessians: BFGS 167

In previous chapters, we have taken a largely *variational* approach to deriving algorithms for computational linear algebra. That is, we define an *objective function*, possibly with constraints, and pose our algorithms as a minimization or maximization problem. A sampling of problems in this domain that we already have considered is listed below:

Problem	**Section**	**Objective**	**Constraints**
Least-squares	3.1.2	$E(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2$	None
Project $\vec{b}$ onto $\vec{a}$	4.4.1	$E(c) = \|c\vec{a} - \vec{b}\|_2^2$	None
Eigenvectors of symmetric A	5.1	$E(\vec{x}) = \vec{x}^\top A\vec{x}$	$\|\vec{x}\|_2 = 1$
Pseudoinverse	6.2.1	$E(\vec{x}) = \|\vec{x}\|_2^2$	$A^\top A\vec{x} = A^\top\vec{b}$
Principal component analysis	6.2.5	$E(C) = \|X - CC^\top X\|_{\text{Fro}}$	$C^\top C = I_{d\times d}$
Broyden step	7.2.2	$E(J_k) = \|J_k - J_{k-1}\|_{\text{Fro}}^2$	$J_k \cdot \Delta\vec{x}_k = \Delta f_k$

The formulation of problems in this fashion is a powerful and general approach. For this reason, it is valuable to design algorithms that function in the absence of a special form for the energy E, in the same way that we developed strategies for finding roots of f without knowing the form of f *a priori* in Chapter 7.

8.1 UNCONSTRAINED OPTIMIZATION: MOTIVATION

In this chapter, we will consider *unconstrained* problems, that is, problems that can be posed as minimizing or maximizing a function $f : \mathbb{R}^n \to \mathbb{R}$ without any constraints on its input $\vec{x}$. It is not difficult to encounter such problems in practice; we list a few examples below.

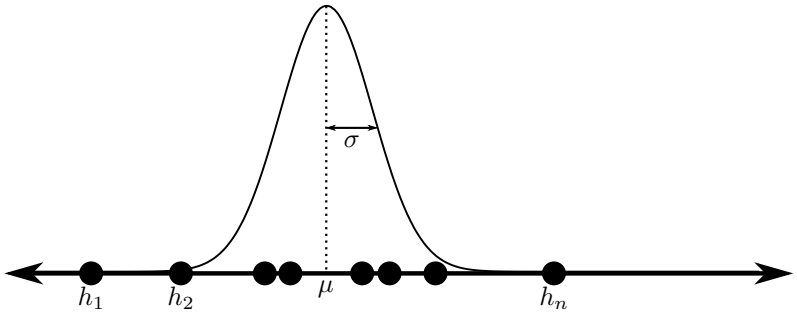


FIGURE 8.1 Illustration for Example 8.2. Given the heights $h_1, h_2, \ldots, h_n$ of students in a class, we we may wish to estimate the mean μ and standard deviation σ of the most likely normal distribution explaining the observed heights.

Example 8.1 (Nonlinear least-squares)**.** Suppose we are given a number of pairs (x_i, y_i) such that $f(x_i) \approx y_i$, and we wish to find the best approximating f within a particular class. For instance, if we expect that f is exponential, we should be able to write $f(x) = ce^{ax}$ for some c and some a; our job is to find the parameters a and c. One strategy we already have developed in Chapter 3 is to minimize the following energy function:

$$E(a, c) = \sum_i (y_i - ce^{ax_i})^2.$$

This form for E is not quadratic in a, so the linear least-squares methods from §3.1.2 do not apply to this minimization problem. Hence, we must find alternative methods to minimize E.

Example 8.2 (Maximum likelihood estimation)**.** In machine learning, the problem of *parameter estimation* involves examining the results of a randomized experiment and trying to summarize them using a probability distribution of a particular form. For example, we might measure the height of every student in a class, yielding a list of heights h_i for each student i. If we have a lot of students, we might model the distribution of student heights using a *normal distribution*:

$$g(h; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(h-\mu)^2/2\sigma^2},$$

where μ is the mean of the distribution and σ is the standard deviation. This notation is illustrated in Figure 8.1.

Under this normal distribution, the likelihood that we observe height h_i for student i is given by $g(h_i; \mu, \sigma)$, and under the (reasonable) assumption that the height of student i is probabilistically independent of that of student j, the probability of observing the entire set of heights observed is given by the product

$$P(\{h_1, \ldots, h_n\}; \mu, \sigma) = \prod_i g(h_i; \mu, \sigma).$$

A common method for estimating the parameters μ and σ of g is to maximize P viewed as a function of μ and σ with $\{h_i\}$ fixed; this is called the *maximum-likelihood*

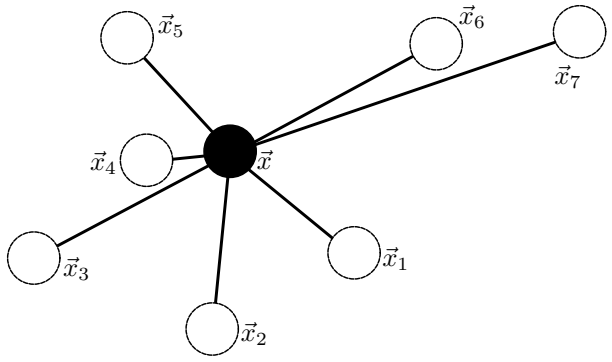


FIGURE 8.2 The geometric median problem seeks a point $\vec{x}$ minimizing the total (non-squared) distance to a set of data points $\vec{x}_1, \ldots, \vec{x}_k$.

estimate of μ and σ. In practice, we usually optimize the *log likelihood* $\ell(\mu, \sigma) \equiv \log P(\{h_1, \ldots, h_n\}; \mu, \sigma)$; this function has the same maxima but enjoys better numerical and mathematical properties.

Example 8.3 (Geometric problems)**.** Many geometry problems encountered in graphics and vision do not reduce to least-squares energies. For instance, suppose we have a number of points $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$. If we wish to *cluster* these points, we might wish to summarize them with a single $\vec{x}$ minimizing:

$$E(\vec{x}) \equiv \sum_i \|\vec{x} - \vec{x}_i\|_2.$$

The $\vec{x}$ minimizing E is known as the *geometric median* of $\{\vec{x}_1, \ldots, \vec{x}_k\}$, as illustrated in Figure 8.2. Since the norm of the difference $\vec{x} - \vec{x}_i$ in E is not squared, the energy is no longer quadratic in the components of $\vec{x}$.

Example 8.4 (Physical equilibria, adapted from [54])**.** Suppose we attach an object to a set of springs; each spring is anchored at point $\vec{x}_i \in \mathbb{R}^3$ and has natural length L_i and constant k_i. In the absence of gravity, if our object is located at position $\vec{p} \in \mathbb{R}^3$, the network of springs has potential energy

$$E(\vec{p}) = \frac{1}{2} \sum_i k_i \left(\|\vec{p} - \vec{x}_i\|_2 - L_i \right)^2.$$

Equilibria of this system are given by minima of E and reflect points $\vec{p}$ at which the spring forces are all balanced. Such systems of equations are used to visualize graphs $G = (V, E)$, by attaching vertices in V with springs for each pair in E.

8.2 OPTIMALITY

Before discussing how to minimize or maximize a function, we should be clear what it is we are looking for; maximizing f is the same as minimizing $-f$, so the minimization problem is sufficient for our consideration. With this goal in mind, for a particular $f : \mathbb{R}^n \to \mathbb{R}$ and

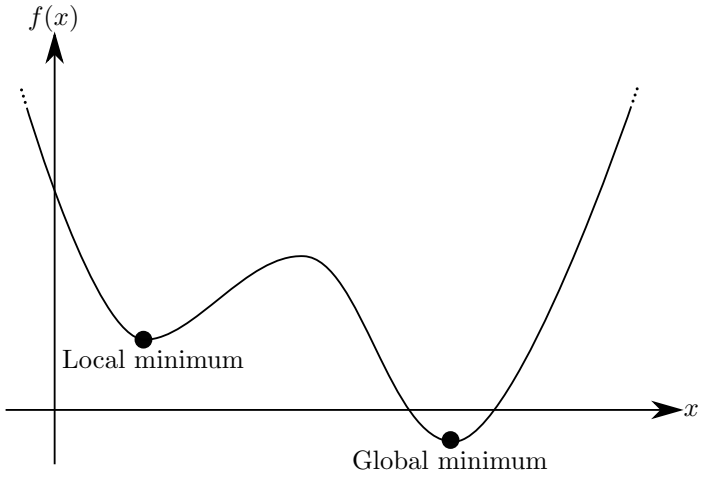


FIGURE 8.3 A function $f(x)$ with two local minima but only one global minimum.

$\vec{x}^* \in \mathbb{R}^n$, we need to derive *optimality conditions* that verify that $\vec{x}^*$ has the lowest possible value $f(\vec{x}^*)$.

Ideally we would like to find *global* minima of f:

Definition 8.1 (Global minimum)**.** The point $\vec{x}^* \in \mathbb{R}^n$ is a *global minimum* of $f : \mathbb{R}^n \to \mathbb{R}$ if $f(\vec{x}^*) \leq f(\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$.

Finding a global minimum of f without any information about the structure of f effectively requires searching in the dark. For instance, suppose an optimization algorithm identifies the left local minimum in the function in Figure 8.3. It is nearly impossible to realize that there is a second, lower minimum by guessing x values—and for all we know, there may be a third even lower minimum of f miles to the right!

Thus, in many cases we are satisfied if we can find a *local* minimum:

Definition 8.2 (Local minimum)**.** The point $\vec{x}^* \in \mathbb{R}^n$ is a *local minimum* of $f : \mathbb{R}^n \to \mathbb{R}$ if there exists some $\varepsilon > 0$ such that $f(\vec{x}^*) \leq f(\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$ satisfying $\|\vec{x} - \vec{x}^*\| < \varepsilon$.

This definition requires that $\vec{x}^*$ attains the smallest value in some *neighborhood* defined by the radius ε. Local optimization algorithms have the severe limitation that they cannot guarantee that they yield the lowest possible value of f, as in Figure 8.3 if the left local minimum is reached. Many strategies, heuristic and otherwise, are applied to explore the landscape of possible $\vec{x}$'s to help gain confidence that a local minimum has the best possible value.

8.2.1 Differential Optimality

A familiar story from single- and multi-variable calculus is that finding potential minima and maxima of a function $f : \mathbb{R}^n \to \mathbb{R}$ is more straightforward when f is differentiable. In this case, the gradient vector $\nabla f = (\partial f/\partial x_1, \ldots, \partial f/\partial x_n)$ at $\vec{x}$ points in the direction moving from $\vec{x}$ in which f increases at the fastest rate; the vector $-\nabla f$ points in the direction of greatest decrease. One way to see this is to approximate that near a point $\vec{x}_0 \in \mathbb{R}^n$, f looks like the linear function

$$f(\vec{x}) \approx f(\vec{x}_0) + \nabla f(\vec{x}_0) \cdot (\vec{x} - \vec{x}_0).$$

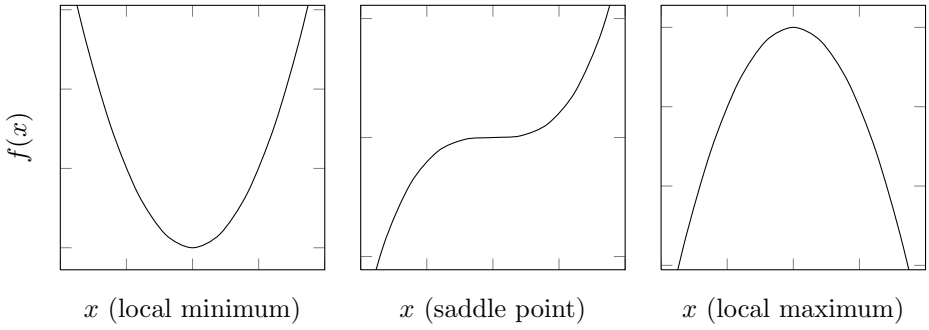


FIGURE 8.4 Different types of critical points.

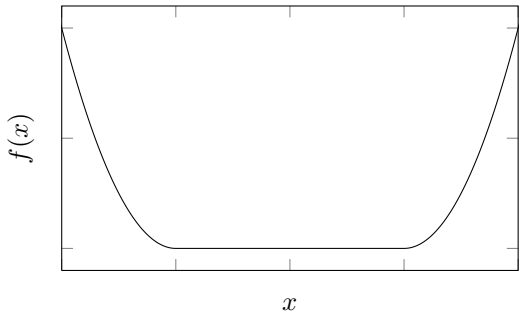


FIGURE 8.5 A function with many stationary points.

If we take $\vec{x} - \vec{x}_0 = \alpha \nabla f(\vec{x}_0)$, then we find

$$f(\vec{x}_0 + \alpha \nabla f(\vec{x}_0)) \approx f(\vec{x}_0) + \alpha \|\nabla f(\vec{x}_0)\|_2^2.$$

The value $\|\nabla f(\vec{x}_0)\|_2^2$ is *always* nonnegative, so the sign of α determines whether f increases or decreases.

It is not difficult to formalize the above argument to show that if $\vec{x}_0$ is a local minimum, then we must have $\nabla f(\vec{x}_0) = \vec{0}$. This condition is *necessary* but not *sufficient*: Maxima and saddle points also have $\nabla f(\vec{x}_0) = \vec{0}$ as illustrated in Figure 8.4. Even so, this observation about minima of differentiable functions yields a high-level approach to optimization:

1. Find points $\vec{x}_i$ satisfying $\nabla f(\vec{x}_i) = \vec{0}$.
2. Check which of these points is a local minimum as opposed to a maximum or saddle point.

Given their role in optimization, we give the points $\vec{x}_i$ a special name:

Definition 8.3 (Stationary point)**.** A *stationary point* of $f : \mathbb{R}^n \to \mathbb{R}$ is a point $\vec{x} \in \mathbb{R}^n$ satisfying $\nabla f(\vec{x}) = \vec{0}$.

Our methods for minimization mostly will find stationary points of f and subsequently eliminate those that are not minima.

It is imperative to keep in mind when we can expect our minimization algorithms to succeed. In most cases, such as those shown in Figure 8.4, the stationary points of f are

isolated, meaning we can write them in a discrete list $\{\vec{x}_0, \vec{x}_1, \ldots\}$. A degenerate case, however, is shown in Figure 8.5; here, an entire interval of values x is composed of stationary points, making it impossible to consider them one at a time. For the most part, we will ignore such issues as poorly-conditioned degenerate cases.

Suppose we identify a point $\vec{x} \in \mathbb{R}$ as a stationary point of f and now wish to check if it is a local minimum. If f is twice-differentiable, we can use its *Hessian* matrix

$$H_f(\vec{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial^2 x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial^2 x_n} \end{pmatrix}.$$

Adding a term to the linearization of f reveals the role of H_f:

$$f(\vec{x}) \approx f(\vec{x}_0) + \nabla f(\vec{x}_0) \cdot (\vec{x} - \vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^\top H_f(\vec{x} - \vec{x}_0).$$

If we substitute a stationary point $\vec{x}^*$, then since the derivative of f equals zero at $\vec{x}^*$ by definition we know

$$f(\vec{x}) \approx f(\vec{x}^*) + \frac{1}{2}(\vec{x} - \vec{x}^*)^\top H_f(\vec{x} - \vec{x}^*).$$

If H_f is positive definite, then this expression shows $f(\vec{x}) \geq f(\vec{x}^*)$, and thus $\vec{x}^*$ is a local minimum. More generally, one of a few situations can occur:

- If H_f is *positive definite*, then $\vec{x}^*$ is a local minimum of f.
- If H_f is *negative definite*, then $\vec{x}^*$ is a local maximum of f.
- If H_f is *indefinite*, then $\vec{x}^*$ is a saddle point of f.
- If H_f is *not invertible*, then oddities such as the function in Figure 8.5 can occur; this includes the case where H_f is *semidefinite*.

Checking if a Hessian matrix is positive definite can be accomplished by checking if its Cholesky factorization exists or—more slowly—by checking that all its eigenvalues are positive. So, when f is sufficiently smooth and the Hessian of f is known, we can check stationary points for optimality using the list above; many optimization algorithms including the ones we will discuss ignore the non-invertible case and notify the user, since it is relatively unlikely.

8.2.2 Alternative Conditions for Optimality

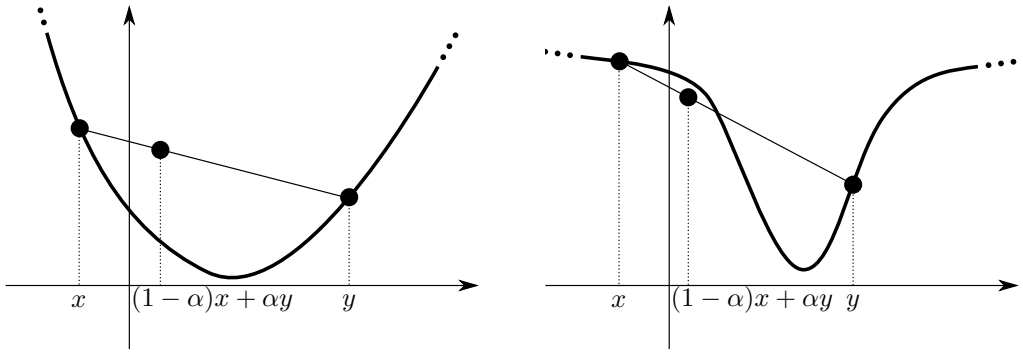
Occasionally, if we know more information about $f : \mathbb{R}^n \to \mathbb{R}$ we can provide optimality conditions that are stronger or easier to check than the ones above. These conditions also can help in the case that f is not differentiable but has other geometric properties that make it possible to find a minimum.

One property of f that has strong implications for optimization is *convexity*, illustrated in Figure 8.6(a):

Definition 8.4 (Convex)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* when for all $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $\alpha \in (0, 1)$ the following relationship holds:

$$f((1-\alpha)\vec{x} + \alpha\vec{y}) \leq (1-\alpha)f(\vec{x}) + \alpha f(\vec{y}).$$

When the inequality is strict (replace $\leq$ with $<$), the function is *strictly convex*.



(a) Convex (b) Quasiconvex

FIGURE 8.6 Convex functions must be bowl-shaped, while quasiconvex functions can have more complicated features.

Convexity implies that if you connect two points in $\mathbb{R}^n$ with a line, the values of f along the line are less than or equal to those you would obtain by linear interpolation.

Convex functions enjoy many strong properties, the most basic of which is the following:

Proposition 8.1. A local minimum of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is necessarily a global minimum.

Proof. Take $\vec{x}$ to be such a local minimum and suppose there exists $\vec{x}^*$ with $f(\vec{x}^*) < f(\vec{x})$. Then, for sufficiently small $\alpha \in (0, 1)$,

$$\begin{aligned} f(\vec{x}) &\leq f(\vec{x} + \alpha(\vec{x}^* - \vec{x})) \text{ since } \vec{x} \text{ is a local minimum} \\ &\leq (1 - \alpha)f(\vec{x}) + \alpha f(\vec{x}^*) \text{ by convexity.} \end{aligned}$$

Moving terms in the inequality $f(\vec{x}) \leq (1 - \alpha)f(\vec{x}) + \alpha f(\vec{x}^*)$ shows $f(\vec{x}) \leq f(\vec{x}^*)$. This contradicts our assumption that $f(\vec{x}^*) < f(\vec{x})$, so $\vec{x}$ must minimize f globally. □

This proposition and related observations show that it is possible to check if you have reached a *global* minimum of a convex function by applying first-order optimality. Thus, it is valuable to check by hand if a function being optimized happens to be convex, a situation occurring surprisingly often in scientific computing; one sufficient condition that can be easier to check when f is twice differentiable is that H_f is positive definite *everywhere.*

Other optimization techniques have guarantees under other assumptions about f. For example, one weaker version of convexity is *quasi*-convexity, achieved when

$$f((1 - \alpha)\vec{x} + \alpha\vec{y}) \leq \max(f(\vec{x}), f(\vec{y})).$$

An example of a quasiconvex function is shown in Figure 8.6(b); although it does not have the characteristic "bowl" shape of a convex function, its local minimizers are necessarily global minimizers.

8.3 ONE-DIMENSIONAL STRATEGIES

As in the last chapter, we will start with optimization of functions $f : \mathbb{R} \to \mathbb{R}$ of one variable and then expand our consideration to more general functions $f : \mathbb{R}^n \to \mathbb{R}$.

8.3.1 Newton's Method

Our principal strategy for minimizing differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ will be to find stationary points $\vec{x}^*$ satisfying $\nabla f(\vec{x}^*) = 0$. Assuming we can check whether stationary points are maxima, minima, or saddle points as a post-processing step, we will focus on the problem of finding the stationary points $\vec{x}^*$.

To this end, suppose $f : \mathbb{R} \to \mathbb{R}$ is twice-differentiable. Then, as in our derivation of Newton's method for root-finding in §7.1.4, we can approximate:

$$f(x) \approx f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2.$$

The approximation on the right hand side is a parabola whose vertex is located at $x_k - {f'(x_k)}/{f''(x_k)}$; we need to include second-order terms since linear functions have no nontrivial minima or maxima.

In reality f is not necessarily a parabola, so minimizing this approximation will not necessarily give a critical point of f directly. So, Newton's method for minimization repeatedly finds the minimum of the parabolic approximation:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}.$$

This technique is easily analyzed given the work we already have put into understanding Newton's method for root-finding in the previous chapter. In particular, an alternative way to derive the iterative formula above comes from root-finding on $f'(x)$, since stationary points satisfy $f'(x) = 0$. Thus, in most cases Newton's method for optimization exhibits quadratic convergence, provided the initial guess x_0 is sufficiently close to x^*.

A natural question is whether the secant method can be applied in an analogous way. Our derivation of Newton's method above finds roots of f', so the secant method could be used to eliminate the evaluation of f'' but not f' from the optimization formula; situations in which we know f' but not f'' are relatively rare. A more suitable parallel is to replace the line segments used to approximate f in the secant method with parabolas. This algorithm, known as *successive parabolic interpolation*, also minimizes a quadratic approximation of f at each iteration, but rather than using $f(x_k)$, $f'(x_k)$, and $f''(x_k)$ to construct the approximation it uses $f(x_k)$, $f(x_{k-1})$, and $f(x_{k-2})$. This technique can converge superlinearly; in practice, however, it can have drawbacks that make other methods discussed in this chapter more preferable. We explore its design in exercise 8.3.

8.3.2 Golden Section Search

Since Newton's method for optimization is so closely linked to root-finding, we might ask whether a similar adaptation can be applied to bisection. Unfortunately, this transition is not obvious. Our motivation for bisection was that it employed only the weakest assumption on f needed to find roots: continuity. The Intermediate Value Theorem does not apply to extrema of a function in any intuitive way, however, so it appears that directly using bisection to minimize a function is not so straightforward.

It is valuable, however, to have at least one minimization algorithm available that does not require differentiability of f as an underlying assumption; after all, there are non-differentiable functions that have clear minima, like $f(x) \equiv |x|$ at $x = 0$. To this end, one alternative assumption might be that f is *unimodular*:

```
function GOLDEN-SECTION-SEARCH(f(x), a, b)
  τ ← ½(√5 − 1)
  x_0 ← a + (1 − τ)(b − a)          ▷ Initial division of interval a < x_0 < x_1 < b
  x_1 ← a + τ(b − a)
  f_0 ← f(x_0)                      ▷ Function values at x_0 and x_1
  f_1 ← f(x_1)
  for k ← 1, 2, 3, ...
    if |b − a| < ε then             ▷ Golden section search converged
      return x* = ½(a + b)
    else if f_0 ≥ f_1 then          ▷ Remove the interval [a, x_0]
      a ← x_0                       ▷ Move left side
      x_0 ← x_1                     ▷ Reuse previous iteration
      f_0 ← f_1
      x_1 ← a + τ(b − a)            ▷ Generate new sample
      f_1 ← f(x_1)
    else if f_1 > f_0 then          ▷ Remove the interval [x_1, b]
      b ← x_1                       ▷ Move right side
      x_1 ← x_0                     ▷ Reuse previous iteration
      f_1 ← f_0
      x_0 ← a + (1 − τ)(b − a)      ▷ Generate new sample
      f_0 ← f(x_0)
```

FIGURE 8.7 The golden section search algorithm finds minima of unimodular functions $f(x)$ on the interval $[a, b]$ even if they are not differentiable.

Definition 8.5 (Unimodular)**.** A function $f : [a, b] \to \mathbb{R}$ is *unimodular* if there exists $x^* \in [a, b]$ such that f is decreasing (or non-increasing) for $x \in [a, x^*]$ and increasing (or non-decreasing) for $x \in [x^*, b]$.

In other words, a unimodular function decreases for some time, and then begins increasing; no localized minima are allowed. Functions like $|x|$ are not differentiable but still are unimodular.

Suppose we have two values x_0 and x_1 such that $a < x_0 < x_1 < b$. We can make two observations that will help us formulate an optimization technique for a unimodular function $f(x)$:

- If $f(x_0) \geq f(x_1)$, then we know that $f(x) \geq f(x_1)$ for all $x \in [a, x_0]$. Thus, the interval $[a, x_0]$ can be discarded in our search for a minimum of f.
- If $f(x_1) \geq f(x_0)$, then we know that $f(x) \geq f(x_0)$ for all $x \in [x_1, b]$, and thus we can discard $[x_1, b]$.

This structure suggests a bisection-like minimization algorithm beginning with the interval $[a, b]$ and iteratively removing pieces according to the rules above.

In such an algorithm, we could remove a *third* of the interval each iteration; this requires two evaluations of f during each iteration at $x_0 = {}^{2a}/_3 + {}^{b}/_3$ and $x_1 = {}^{a}/_3 + {}^{2b}/_3$. If evaluating f is expensive, however, we may wish to reuse information from previous iterations to eliminate one of those two evaluations.

For now take $a = 0$ and $b = 1$; the strategies we derive below eventually will work more generally by shifting and scaling. In the absence of more information about f, we might as

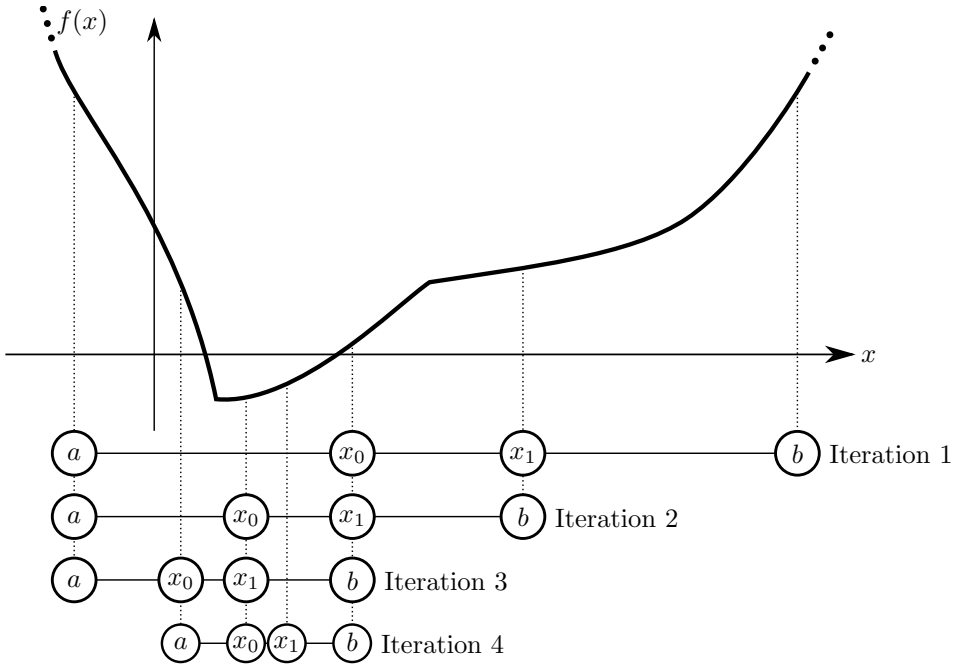


FIGURE 8.8 Iterations of golden section search on unimodular $f(x)$ shrink the interval $[a, b]$ by eliminating the left segment $[a, x_0]$ or the right segment $[x_1, b]$; each iteration reuses either $f(x_0)$ or $f(x_1)$ via the construction in §8.3.2. In this illustration, each horizontal line represents an iteration of golden section search, with the values a, x_0, x_1, and b labeled in the circles.

well make a symmetric choice $x_0 = \alpha$ and $x_1 = 1 - \alpha$ for some $\alpha \in (0, 1/2)$. Suppose our iteration removes the rightmost interval $[x_1, b]$. Then, the search interval becomes $[0, 1-\alpha]$, and we know $f(\alpha)$ from the previous iteration. The next iteration will divide $[0, 1-\alpha]$ such that $x_0 = \alpha(1-\alpha)$ and $x_1 = (1-\alpha)^2$. If we wish to reuse $f(\alpha)$ from the previous iteration, we could set $(1-\alpha)^2 = \alpha$, yielding:

$$\alpha = \frac{1}{2}(3 - \sqrt{5})$$
$$1 - \alpha = \frac{1}{2}(\sqrt{5} - 1)$$

The value of $1-\alpha \equiv \tau$ above is the *golden ratio*! It allows for the reuse of one of the function evaluations from the previous iterations; a symmetric argument shows that the same choice of α works if we had removed the left interval instead of the right one.

The *golden section search* algorithm, documented in Figure 8.7 and illustrated in Figure 8.8, makes use of this construction to minimize a unimodular function $f(x)$ on the interval $[a, b]$ via subdivision with one evaluation of $f(x)$ per iteration. This algorithm converges unconditionally and linearly, since a fraction α of the interval $[a, b]$ bracketing the minimum is removed in each iteration.

When f is not globally unimodal, golden section search does not apply unless we can find some $[a, b]$ such that f is unimodal on that interval. In some cases, $[a, b]$ can be guessed by attempting to bracket a local minimum of f. For example, [96] suggests stepping farther

```
function GRADIENT-DESCENT(f(x⃗), x⃗_0)
  x⃗ ← x⃗_0
  for k ← 1, 2, 3, ...
    DEFINE-FUNCTION(g(t) ≡ f(x⃗ − t∇f(x⃗)))
    t* ← LINE-SEARCH(g(t), t ≥ 0)
    x⃗ ← x⃗ − t∇f(x⃗)            ▷ Update estimate of minimum
    if ‖∇f(x⃗)‖ < ε then
      return x* = x⃗
```

FIGURE 8.9 The gradient descent algorithm iteratively minimizes $f : \mathbb{R}^n \to \mathbb{R}$ by solving one-dimensional minimizations through the gradient direction. LINE-SEARCH can be one of the methods from §8.3 for minimization in one dimension. In faster, more advanced techniques, this method can find suboptimal $t^* > 0$ that still decreases $g(t)$ sufficiently to make sure the optimization does not get stuck.

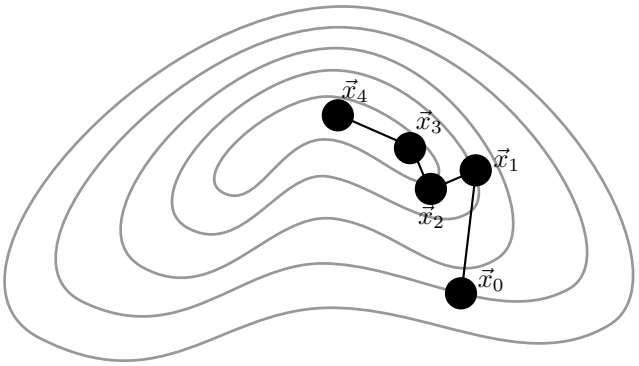


FIGURE 8.10 Gradient descent on a function $f : \mathbb{R}^2 \to \mathbb{R}$, whose level sets are shown in gray. The gradient $\nabla f(\vec{x})$ points perpendicular to the level sets of f, as in Figure 0.6; gradient descent iteratively minimizes f along the line through this direction.

and farther away from some starting point $x_0 \in \mathbb{R}$, moving downhill from $f(x_0)$ until f increases again, suggesting the presence of a local minimum.

8.4 MULTIVARIABLE STRATEGIES

We continue to parallel our discussion of root-finding by expanding our discussion from single-variable to multivariable problems. As with root-finding, multivariable optimization problems are considerably more difficult than problems in a single variable, but they appear so many times in practice that they are worth careful consideration.

Here, we will consider only the case that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable. Optimization methods more similar to golden section search for non-differentiable functions are less common and are difficult to formulate. See e.g. [69, 15] for consideration of non-differentiable optimization, subgradients, and related concepts.

8.4.1 Gradient Descent

From our previous discussion, $\nabla f(\vec{x})$ points in the direction of "steepest ascent" of f at $\vec{x}$ and that $-\nabla f(\vec{x})$ is the direction of "steepest descent." If nothing else, this definition suggests that when $\nabla f(\vec{x}) \neq \vec{0}$, for small $\alpha > 0$ we must have

$$f(\vec{x} - \alpha \nabla f(\vec{x})) \leq f(\vec{x}).$$

Suppose our current estimate of the location of the minimum of f is $\vec{x}_k$. Then, a reasonable iterative minimization strategy should seek the next iterate $\vec{x}_{k+1}$ so that $f(\vec{x}_{k+1}) < f(\vec{x}_k)$. Since we do not expect to find a global minimum in one shot, we can make restricting assumptions to simplify the search for the next iterate. One typical way to simplify the search for $\vec{x}_{k+1}$ is to use a one-variable algorithm from §8.3 on f after restricting it to a line through $\vec{x}_k$; once we have settled on $\vec{x}_{k+1}$ we will choose a new line and optimize again.

Consider the function $g_k(t) \equiv f(\vec{x}_k - t\nabla f(\vec{x}_k))$, which restricts f to the line through $\vec{x}_k$ parallel to $\nabla f(\vec{x}_k)$. We know that when $\nabla f(\vec{x}_k) \neq \vec{0}$, substituting small $t > 0$ will yield a decrease in f. Hence, this is a reasonable direction for a restricted search along a line for a new iterate. The *gradient descent* algorithm shown in Figure 8.9 and illustrated in Figure 8.10 iteratively solves these one-dimensional problems to improve $\vec{x}_k$.

Each iteration of gradient descent decreases $f(\vec{x}_k)$, so these values converge assuming they are bounded below. The algorithm only terminates when $\nabla f(\vec{x}_k) \approx \vec{0}$, showing that gradient descent must at least reach a local minimum; convergence is slow for most functions f, however. The line search process can be replaced by a method that decreases the objective a non-negligible if suboptimal amount, although it is more difficult to guarantee convergence in this case.

8.4.2 Newton's Method in Multiple Variables

Paralleling our derivation of the single-variable case in §8.3.1, we can write a Taylor series approximation of $f : \mathbb{R}^n \to \mathbb{R}$ using its Hessian matrix H_f:

$$f(\vec{x}) \approx f(\vec{x}_k) + \nabla f(\vec{x}_k)^\top \cdot (\vec{x} - \vec{x}_k) + \frac{1}{2}(\vec{x} - \vec{x}_k)^\top \cdot H_f(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k).$$

Differentiating with respect to $\vec{x}$ and setting the result equal to zero yields the following iterative scheme:

$$\vec{x}_{k+1} = \vec{x}_k - [H_f(\vec{x}_k)]^{-1} \nabla f(\vec{x}_k).$$

This expression is a generalization of the analogous method in §8.3.1, and once again it converges quadratically when $\vec{x}_0$ is near a minimum.

Newton's method can be more efficient than gradient descent depending on the optimization objective f since it makes use of both first- and second-order information about f. It is intuitive why Newton's method converges quickly when it is near an optimum. In particular, gradient descent has no knowledge of H_f; it proceeds analogously to walking downhill by looking only at your feet. By using H_f, Newton's method has a larger picture of the shape of f nearby.

Each iteration of gradient descent potentially requires many evaluations of f during the line search procedure. On the other hand, we must evaluate and invert the Hessian H_f during each iteration of Newton's method. These factors do not affect the number of iterations but do affect runtime; this tradeoff does not appear in traditional analysis.

When H_f is nearly singular, Newton's method potentially can take very large steps

away from the current estimate of the minimum. These large steps are a good idea if the second-order approximation of f is accurate, but as the step becomes large the quality of this approximation can degenerate. One way to take more conservative steps is to "dampen" the change in $\vec{x}$ using a small multiplier $\gamma > 0$:

$$\vec{x}_{k+1} = \vec{x}_k - \gamma[H_f(\vec{x}_k)]^{-1}\nabla f(\vec{x}_k)$$

A more expensive but safer strategy is to do line search from $\vec{x}_k$ along the direction $-[H_f(\vec{x}_k)]^{-1}\nabla f(\vec{x}_k)$.

When H_f is not positive definite the objective locally might look like a saddle or peak rather than a bowl. In this case, jumping to an approximate stationary point might not make sense. Thus, adaptive techniques might check if H_f is positive definite before applying a Newton step; if it is not positive definite, the methods revert to gradient descent to find a better approximation of the minimum. Alternatively, they can modify H_f, for example by projecting onto the closest positive definite matrix.

8.4.3 Optimization without Hessians: BFGS

Newton's method can be difficult to apply to complicated functions $f : \mathbb{R}^n \to \mathbb{R}$. The second derivative of f might be considerably more expensive to evaluate than f or even ∇f, and H_f changes with each iteration, eliminating the possibility of pre-factorization. Additionally, H_f has size $n \times n$, so storing H_f requires $O(n^2)$ space, which might be too large. Newton's method deals with *approximations* of f in each iteration anyway, so we might attempt to formulate less expensive second-order approximations that still outperform gradient descent.

As in our discussion of root-finding in §7.2.2, techniques for minimization that imitate Newton's method but use approximate derivatives are called *quasi-Newton methods*. Often they can have similarly strong convergence properties without the need for explicit re-evaluation and even inversion of the Hessian at each iteration. Here, we will follow the development of [85] to motivate one modern technique for quasi-Newton optimization.

Suppose we wish to minimize $f : \mathbb{R}^n \to \mathbb{R}$ using an iterative scheme. Near the current estimate $\vec{x}_k$ of the root, we might estimate f with a quadratic model:

$$f(\vec{x}_k + \delta\vec{x}) \approx f(\vec{x}_k) + \nabla f(\vec{x}_k) \cdot \delta\vec{x} + \frac{1}{2}(\delta\vec{x})^\top B_k(\delta\vec{x}).$$

Here, we require that our approximation agrees with f to first order at $\vec{x}_k$, but we will allow our estimate of the Hessian B_k to differ from the actual Hessian of f.

Slightly generalizing Newton's method in §8.4.2, this quadratic approximation is minimized by taking $\delta\vec{x} = -B_k^{-1}\nabla f(\vec{x}_k)$. In case $\|\delta\vec{x}\|_2$ is large and we do not wish to take such a large step, we will allow ourselves to scale this difference by a step size α_k, yielding

$$\vec{x}_{k+1} = \vec{x}_k - \alpha_k B_k^{-1}\nabla f(\vec{x}_k).$$

Our goal is estimate B_{k+1} by updating B_k, so that we can repeat this process.

The Hessian of f is nothing more than the derivative of ∇f, so like Broyden's method we can use previous iterates to impose a secant-style condition on B_{k+1}:

$$B_{k+1}(\vec{x}_{k+1} - \vec{x}_k) = \nabla f(\vec{x}_{k+1}) - \nabla f(\vec{x}_k).$$

For convenience of notation, we will define $\vec{s}_k \equiv \vec{x}_{k+1} - \vec{x}_k$ and $\vec{y}_k \equiv \nabla f(\vec{x}_{k+1}) - \nabla f(\vec{x}_k)$, yielding an equivalent condition $B_{k+1}\vec{s}_k = \vec{y}_k$.

Given the optimization at hand, we wish for B_k to have two properties:

```
function BFGS(f(x⃗), x⃗_0)
  H ← I_{n×n}
  x⃗ ← x⃗_0
  for k ← 1, 2, 3, ...
    if ‖∇f(x⃗)‖ < ε then
      return x* = x⃗
    p⃗ ← −H_k ∇f(x⃗)                          ▷ Next search direction
    α ← Compute-Alpha(f, p⃗, x⃗, y⃗)           ▷ Satisfy positive definite condition
    s⃗ ← αp⃗                                  ▷ Displacement of x⃗
    x⃗ ← x⃗ + s⃗                              ▷ Update estimate
    y⃗ ← ∇f(x⃗ + s⃗) − ∇f(x⃗)                  ▷ Change in gradient
    ρ ← 1/(y⃗·s⃗)          ▷ Apply BFGS update to inverse Hessian approximation
    H ← (I_{n×n} − ρs⃗y⃗^⊤)H(I_{n×n} − ρy⃗s⃗^⊤) + ρs⃗s⃗^⊤
```

FIGURE 8.11 The BFGS algorithm for finding a local minimum of differentiable $f(\vec{x})$ without its Hessian. The function Compute-Alpha finds large $\alpha > 0$ satisfying $\vec{y} \cdot \vec{s} > 0$, where $\vec{y} = \nabla f(\vec{x} + \vec{s}) - \nabla f(\vec{x})$ and $\vec{s} = \alpha\vec{p}$.

- B_k should be a symmetric matrix, like the Hessian H_f.
- B_k should be positive (semi-)definite, so that we are seeking minima of f rather than maxima or saddle points.

The symmetry condition eliminates the possibility of using the Broyden estimate we developed in the previous chapter.

The positive definite constraint implicitly puts a condition on the relationship between $\vec{s}_k$ and $\vec{y}_k$. In particular, premultiplying the relationship $B_{k+1}\vec{s}_k = \vec{y}_k$ by $\vec{s}_k^\top$ shows $\vec{s}_k^\top B_{k+1}\vec{s}_k = \vec{s}_k^\top \vec{y}_k$. For B_{k+1} to be positive definite, we must then have $\vec{s}_k \cdot \vec{y}_k > 0$. This observation can guide our choice of α_k; it must hold for sufficiently small $\alpha_k > 0$.

Assume that $\vec{s}_k$ and $\vec{y}_k$ satisfy the positive definite compatibility condition. Then, we can write down a Broyden-style optimization problem leading to a possible approximation B_{k+1}:

$$\begin{array}{rl} \text{minimize}_{B_{k+1}} & \|B_{k+1} - B_k\| \\ \text{such that} & B_{k+1}^\top = B_{k+1} \\ & B_{k+1}\vec{s}_k = \vec{y}_k \end{array}$$

For appropriate choice of norms $\|\cdot\|$, this optimization yields the well-known DFP (Davidon-Fletcher-Powell) iterative scheme.

Rather than working out the details of the DFP scheme, we move on to a more popular method known as the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm, in Figure 8.11. Ignoring our choice of α_k for now, our second-order approximation was minimized by taking $\delta\vec{x} = -B_k^{-1}\nabla f(\vec{x}_k)$. Thus, in the end the behavior of our iterative scheme is dictated by the *inverse* matrix B_k^{-1}. Asking that $\|B_{k+1} - B_k\|$ is small can still imply relatively bad differences between the action of B_k^{-1} and that of B_{k+1}^{-1}!

With this observation in mind, the BFGS scheme makes a small alteration to the above derivation. Rather than computing B_k at each iteration, we can compute its inverse $H_k \equiv B_k^{-1}$ directly.* Now our condition $B_{k+1}\vec{s}_k = \vec{y}_k$ gets reversed to $\vec{s}_k = H_{k+1}\vec{y}_k$; the condition

*We choose to use standard notation for BFGS in this section, but a common point of the confusion

that B_k is symmetric is the same as the condition that H_k is symmetric. We solve an optimization problem

$$\begin{array}{rl} \text{minimize}_{H_{k+1}} & \|H_{k+1} - H_k\| \\ \text{such that} & H_{k+1}^\top = H_{k+1} \\ & \vec{s}_k = H_{k+1}\vec{y}_k. \end{array}$$

This construction has the nice side benefit of not requiring matrix inversion to compute $\delta\vec{x} = -H_k \nabla f(\vec{x}_k)$.

To derive a formula for H_{k+1}, we must decide on a matrix norm $\|\cdot\|$. As with our previous discussion, the *Frobenius* norm looks closest to least-squares optimization, making it likely we can generate a closed-form expression for H_{k+1} rather than having to solve the minimization above as a subroutine of BFGS optimization.

The Frobenius norm, however, has one serious drawback for Hessian matrices. The Hessian matrix has entries $(H_f)_{ij} = {}^{\partial f_i}/_{\partial x_j}$. Often the quantities x_i for different i can have different *units*; e.g. consider maximizing the profit (in dollars) made by selling a cheeseburger of radius r (in inches) and price p (in dollars), leading to $f : (\text{inches}, \text{dollars}) \to \text{dollars}$. Squaring these different quantities and adding them up does not make sense.

Suppose we find a symmetric positive definite matrix W so that $W\vec{s}_k = \vec{y}_k$; we will check in the exercises that such a matrix exists. Such a matrix takes the units of $\vec{s}_k = \vec{x}_{k+1} - \vec{x}_k$ to those of $\vec{y}_k = \nabla f(\vec{x}_{k+1}) - \nabla f(\vec{x}_k)$. Taking inspiration from our expression $\|A\|_{\text{Fro}}^2 = \text{Tr}(A^\top A)$, we can define a *weighted Frobenius norm* of a matrix A as

$$\|A\|_W^2 \equiv \text{Tr}(A^\top W^\top A W).$$

Unlike the Frobenius norm of H_{k+1}, this expression has consistent units when applied to our optimization for H_{k+1}. When both W and A are symmetric with columns $\vec{w}_i$ and $\vec{a}_i$, resp., expanding the expression above shows:

$$\|A\|_W^2 = \sum_{ij} (\vec{w}_i \cdot \vec{a}_j)(\vec{w}_j \cdot \vec{a}_i).$$

This choice of norm combined with the choice of W yields a particularly clean formula for H_{k+1} given H_k, $\vec{s}_k$, and $\vec{y}_k$:

$$H_{k+1} = (I_{n\times n} - \rho_k \vec{s}_k \vec{y}_k^\top) H_k (I_{n\times n} - \rho_k \vec{y}_k \vec{s}_k^\top) + \rho_k \vec{s}_k \vec{s}_k^\top,$$

where $\rho_k \equiv {}^1/_{\vec{y}_k \cdot \vec{s}_k}$. We show in the Appendix to this chapter how to derive this formula; the proof requires a number of algebraic steps but conceptually is no more difficult than direct application of Lagrange multipliers for constrained optimization (see Theorem 0.1).

The BFGS algorithm avoids the need to compute and invert a Hessian matrix for f, but it still requires $O(n^2)$ storage for H_k. The L-BFGS ("Limited-Memory BFGS") variant avoids this issue by keeping a limited history of vectors $\vec{y}_k$ and $\vec{s}_k$ and using these to apply H_k by expanding its formula recursively. L-BFGS can have *better* numerical properties despite its compact use of space, since old vectors $\vec{y}_k$ and $\vec{s}_k$ may no longer be relevant and *should* be ignored. Exercise 8.10 derives this technique.

8.5 EXERCISES

8.1 Show that $f(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2$ is a convex function.

is that H now represents an approximate *inverse* Hessian. This is the **not** the same as the Hessian H_f in §8.4.2 and elsewhere.

8.2 Some observations about convex and quasiconvex functions:

(a) Show that every convex function is quasiconvex, but that some quasiconvex functions are not convex.

(b) Show that any local minimum of a continuous, strictly quasiconvex function $f : \mathbb{R}^n \to \mathbb{R}$ is also a global minimum of f. Here, *strict* quasiconvexity replaces the $\leq$ in the definition of quasiconvex functions with $<$.

(c) Show that the sum of two convex functions is convex, but give a counterexample showing that the sum of two quasiconvex functions may not be quasiconvex.

(d) Suppose $f(x)$ and $g(x)$ are quasiconvex. Show that $h(x) = \max(f(x), g(x))$ is quasiconvex.

8.3 In §8.3.1, we suggested the possibility of using parabolas rather than secants to minimize a function $f : \mathbb{R} \to \mathbb{R}$ without knowing any of its derivatives. Here, we outline its design:

(a) Suppose we are given three points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ with distinct x values. Show that the vertex of the parabola $y = ax^2 + bx + c$ through these points is given by:

$$x = x_2 - \frac{(x_2 - x_1)^2(y_2 - y_3) - (x_2 - x_3)^2(y_2 - y_1)}{2(x_2 - x_1)(y_2 - y_3) - (x_2 - x_3)(y_2 - y_1)}$$

(b) Use this formula to propose an iterative technique for minimizing a function of one variable without using any of its derivatives.

(c) What happens when the three points in 8.3a are collinear? Does this suggest a failure mode of successive parabolic interpolation?

(d) Does the formula in 8.3a distinguish between maxima and minima of parabolas? Does this suggest a second failure mode?

8.4 Show that a strictly convex function $f : [a, b] \to \mathbb{R}$ is unimodular.

8.5 We might ask how well we can expect methods like golden section search can work in the presence of finite precision arithmetic. We step through a few analytical steps from [96]:

(a) Suppose we have bracketed a local minimum of differentiable $f(x)$ in a small interval. Justify the following approximation in this interval:

$$f(x) \approx f(b) + \frac{1}{2}f''(b)(x - b)^2$$

(b) Suppose we wish to refine the interval containing the minimum until the second term in this approximation is negligible. Show that if we wish to upper-bound the ratio of the two terms in 8.5a by ε, we should enforce

$$|x - b| < \sqrt{\frac{2\varepsilon|f(b)|}{|f''(b)|}}.$$

(c) By taking ε to be machine precision as in §1.1.2, conclude that the size of the interval in which $f(x)$ and $f(b)$ are indistinguishable numerically grows like $\sqrt{\varepsilon}$. Based on this observation, can golden section search bracket a root within machine precision?
Hint: For small $\varepsilon > 0$, $\sqrt{\varepsilon} \gg \varepsilon$.

8.6[DH] For a convex function $f : U \to \mathbb{R}^n$, where $U \subseteq \mathbb{R}^n$ is convex and open, define a *subgradient* of f at $\vec{x}_0 \in U$ to be any vector $\vec{s} \in \mathbb{R}^n$ such that

$$f(\vec{x}) - f(\vec{x}_0) \geq \vec{s} \cdot (\vec{x} - \vec{x}_0),$$

for any $\vec{x} \in U$ [107]. The subgradient is a plausible choice for generalizing the notion of a gradient at a point where f is not differentiable. The *subdifferential* $\partial f(\vec{x}_0)$ is the set of all subgradients of f at $\vec{x}_0$.

For the remainder of this question, assume that f is convex and continuous:

(a) What is $\partial f(0)$ for the function $f(x) = |x|$?

(b) Suppose we wish to minimize (convex and continuous) $f : \mathbb{R}^n \to \mathbb{R}$, which may not be differentiable everywhere. Propose an optimality condition involving subdifferentials for a point $\vec{x}^*$ to be a minimizer of f. Show that your condition holds if and only if $\vec{x}^*$ globally minimizes f.

(c) The *subgradient method* extends gradient descent to a wider class of functions. Analogously to gradient descent, the subgradient method performs the iteration

$$\vec{x}_{k+1} \equiv \vec{x}_k - \alpha_{k+1}\vec{g}_k,$$

where α_{k+1} is a step size and g_k is *any* subgradient of f at $\vec{x}_k$. We do not necessarily decrease f in each iteration, so instead we keep track of the best iterate we have seen so far, $\vec{x}_k^{\text{best}}$.
In the following parts, assume that we take α to be a constant step size, that f is Lipschitz continuous with constant $C > 0$, and that $\|\vec{x}_1 - \vec{x}^*\| \leq B$ for some $B > 0$; this third assumption implies that the error of our second iterate is bounded. Under these assumptions we will show that

$$\lim_{k\to\infty} f(\vec{x}_k^{\text{best}}) \leq f(\vec{x}^*) + \frac{C^2}{2}\alpha,$$

a bound characterizing convergence of the subgradient method.

(i) Derive an upper bound for the error of $\vec{x}_{k+1}$ in terms of the error of $\vec{x}_k$.
Hint: It suffices to consider the square of each error. Combine the definitions of the subgradient itself as well as the iterative subgradient optimization method.

(ii) By recursively applying the result from part 8.6(c)i, provide an upper bound for the error of $\vec{x}_{k+1}$ in terms of the error of $\vec{x}_1$.
Hint: Again, consider squares of the errors.

(iii) Incorporate $f(\vec{x}_k^{\text{best}})$ and the given scalar bounds into your result and take the limit as $k \to \infty$ to obtain the desired conclusion.

(iv) In practice, rather than keeping α constant we should take $\alpha \to 0$ to find $\vec{x}^*$ without the $C^2\alpha/2$ error term. We must choose α to decrease quickly enough

that this term disappears, but slowly enough that the method converges to the minimizer of f (taking $\alpha \equiv 0$ will never find the minimum!). What is the convergence rate of subgradient descent if we choose $\alpha = {}^{B}/_{C\sqrt{k}}$? *Note:* This convergence rate is optimal for subgradient descent.

8.7[SC] This problem will demonstrate how to project a Hessian onto the nearest positive definite matrix. Some optimization techniques use this operation to avoid attempting to minimize in directions where a function is not bowl-shaped.

(a) Suppose $M, U \in \mathbb{R}^{n\times n}$, where M is symmetric and U is orthogonal. Show that $\|UMU^\top\|_{\text{Fro}} = \|M\|_{\text{Fro}}$.

(b) Decompose $M = Q\Lambda Q^\top$, where Λ is a diagonal matrix of eigenvalues and Q is an orthogonal matrix of eigenvectors. Using the result of the previous part, explain how the positive semidefinite matrix $\bar{M}$ closest to M with respect to the Frobenius norm can be constructed by clamping the negative eigenvalues in Λ to zero.

8.8 Our derivation of the BFGS algorithm in §8.4.3 depended on the existence of a symmetric positive definite matrix W satisfying $W\vec{s}_k = \vec{y}_k$. Show that one such matrix is $W \equiv \bar{G}_k^{-1}$, where $\bar{G}_k$ is the *average Hessian* [85]:

$$\bar{G}_k \equiv \int_0^1 H_f(\vec{x}_k + \tau\vec{s}_k)\, d\tau.$$

Do we ever have to compute W in the course of running BFGS?

8.9 Derive an explicit update formula for obtaining B_{k+1} from B_k in the Davidon-Fletcher-Powell scheme mentioned in §8.4.3. Use the $\|\cdot\|_W$ norm introduced in the derivation of BFGS.

8.10 The matrix H used in the BFGS algorithm generally is dense, requiring $O(n^2)$ storage for $f : \mathbb{R}^n \to \mathbb{R}$. This scaling may be infeasible for large n.

(a) Provide an alternative approach to storing H requiring $O(nk)$ storage in iteration k of BFGS.
Hint: Your algorithm may have to "remember" data from previous iterations.

(b) If we need to run for many iterations, the storage from the previous part can exceed the $O(n^2)$ limit we were attempting to avoid. Propose an approximation to H that uses no more than $O(nk_{\max})$ storage, for a user-specified constant $k_{\max}$.

8.11 The BFGS and DFP algorithms update (inverse) Hessian approximations using matrices of rank two. For simplicity, the *symmetric-rank-1* (SR1) update restricts changes to be rank one instead [85].

(a) Suppose $B_{k+1} = B_k + \sigma\vec{v}\vec{v}^\top$, where $|\sigma| = 1$ and $\vec{y}_k = B_{k+1}\vec{s}_k$. Show that under these conditions we must have:

$$B_{k+1} = B_k + \frac{(\vec{y}_k - B_k\vec{s}_k)(\vec{y}_k - B_k\vec{s}_k)^\top}{(\vec{y}_k - B_k\vec{s}_k)^\top\vec{s}_k}.$$

(b) Suppose $H_k \equiv B_k^{-1}$. Show that H_k can be updated as

$$H_{k+1} = H_k + \frac{(\vec{s}_k - H_k\vec{y}_k)(\vec{s}_k - H_k\vec{y}_k)^\top}{(\vec{s}_k - H_k\vec{y}_k)^\top \vec{y}_k}.$$

Hint: Use the result of problem 7.6.

8.12 Here we examine some changes to the gradient descent algorithm for unconstrained optimization on a function f.

(a) In machine learning, the *stochastic gradient descent* algorithm can be used to optimize many common objective functions:

(i) Give an example of a practical optimization problem with an objective taking the form $f(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} g(\vec{x}_i - \vec{x})$ for some function $g : \mathbb{R}^n \to \mathbb{R}$.

(ii) Propose a randomized approximation of ∇f summing no more than k terms (for some $k \ll N$) assuming the $\vec{x}_i$'s are similar to one another. Discuss advantages and drawbacks of using such an approximation.

(b) The "line search" part of gradient descent must be considered carefully:

(i) Suppose an iterative optimization routine gives a sequence of estimates $\vec{x}_1, \vec{x}_2, \ldots$ of the position $\vec{x}^*$ of the minimum of f. Is it enough to assume $f(\vec{x}_k) < f(\vec{x}_{k-1})$ to guarantee that the $\vec{x}_k$'s converge to a local minimum? Why?

(ii) Suppose we run gradient descent. If we suppose $f(\vec{x}) \geq 0$ for all $\vec{x}$ and that we are able to find t^* exactly in each iteration, show that $f(\vec{x}_k)$ converges as $k \to \infty$.

(iii) Explain how the optimization in 8.12(b)ii for t^* can be overkill. In particular, explain how the Wolfe conditions (you will have to look these up!) relax the assumption that we can find t^*.

8.13 Sometimes we are greedy and wish to optimize multiple objectives simultaneously. For example, we might want to fire a rocket to reach an optimal point in time *and* space. It may not be possible to do this simultaneously, but some theories attempt to reconcile multiple optimization objectives.

Suppose we are given functions $f_1(\vec{x}), f_2(\vec{x}), \ldots, f_k(\vec{x})$. A point $\vec{x}$ is said to *Pareto dominate* another point $\vec{y}$ if $f_i(\vec{x}) \leq f_i(\vec{y})$ for all i and $f_j(\vec{x}) < f_j(\vec{y})$ for some $j \in \{1, \ldots, k\}$. A point $\vec{x}^*$ is *Pareto optimal* if it is not dominated by any point $\vec{y}$. Assume $f_1, \ldots, f_k$ are convex and in particular have unique minimizers.

(a) Show that the set of Pareto optimal points is nonempty in this case.

(b) Suppose $\sum_i \gamma_i = 1$ and $\gamma_i > 0$ for all i. Show that the minimizer $\vec{x}^*$ of $g(\vec{x}) \equiv \sum_i \gamma_i f_i(\vec{x})$ is Pareto optimal.
Note: One strategy for multi-objective optimization is to promote $\vec{\gamma}$ to a variable with constraints $\vec{\gamma} \geq \vec{0}$ and $\sum_i \gamma_i = 1$.

(c) Suppose $\vec{x}_i^*$ minimizes $f_i(\vec{x})$ over all possible $\vec{x}$. Write vector $\vec{z} \in \mathbb{R}^k$ with components $z_i = f_i(\vec{x}_i^*)$. Show that the minimizer $\vec{x}^*$ of $h(\vec{x}) \equiv \sum_i (f_i(\vec{x}) - z_i)^2$ is Pareto optimal.
Note: This part and the previous part represent two possible *scalarizations* of the multi-objective optimization problem by reducing to a single scalar objective.

8.6 APPENDIX: DERIVATION OF BFGS UPDATE

In this optional appendix, we derive in detail the BFGS update from §8.4.3.* Our optimization for H_{k+1} has the following Lagrange multiplier expression (for ease of notation we take $H_{k+1} \equiv H$ and $H_k = H^*$):

$$\begin{aligned}\Lambda &\equiv \sum_{ij}(\vec{w}_i \cdot (\vec{h}_j - \vec{h}_j^*))(\vec{w}_j \cdot (\vec{h}_i - \vec{h}_i^*)) - \sum_{i<j}\alpha_{ij}(H_{ij} - H_{ji}) - \vec{\lambda}^\top(H\vec{y}_k - \vec{s}_k) \\ &= \sum_{ij}(\vec{w}_i \cdot (\vec{h}_j - \vec{h}_j^*))(\vec{w}_j \cdot (\vec{h}_i - \vec{h}_i^*)) - \sum_{ij}\alpha_{ij}H_{ij} - \vec{\lambda}^\top(H\vec{y}_k - \vec{s}_k) \text{ if we assume } \alpha_{ij} = -\alpha_{ji}\end{aligned}$$

Taking derivatives to find critical points shows (for $\vec{y} \equiv \vec{y}_k, \vec{s} \equiv \vec{s}_k$):

$$\begin{aligned}0 = \frac{\partial\Lambda}{\partial H_{ij}} &= \sum_\ell 2w_{i\ell}(\vec{w}_j \cdot (\vec{h}_\ell - \vec{h}_\ell^*)) - \alpha_{ij} - \lambda_i y_j \\ &= 2\sum_\ell w_{i\ell}(W^\top(H - H^*))_{j\ell} - \alpha_{ij} - \lambda_i y_j \\ &= 2\sum_\ell (W^\top(H - H^*))_{j\ell}w_{\ell i} - \alpha_{ij} - \lambda_i y_j \text{ by symmetry of } W \\ &= 2(W^\top(H - H^*)W)_{ji} - \alpha_{ij} - \lambda_i y_j \\ &= 2(W(H - H^*)W)_{ij} - \alpha_{ij} - \lambda_i y_j \text{ by symmetry of } W \text{ and } H\end{aligned}$$

So, in matrix form we have the following list of facts:

$$\begin{aligned}0 &= 2W(H - H^*)W - A - \vec{\lambda}\vec{y}^\top, \text{ where } A_{ij} = \alpha_{ij} \\ A^\top &= -A, W^\top = W, H^\top = H, (H^*)^\top = H^* \\ H\vec{y} &= \vec{s}, W\vec{s} = \vec{y}\end{aligned}$$

We can achieve a pair of relationships using transposition combined with symmetry of H and W and asymmetry of A:

$$\begin{aligned}0 &= 2W(H - H^*)W - A - \vec{\lambda}\vec{y}^\top \\ 0 &= 2W(H - H^*)W + A - \vec{y}\vec{\lambda}^\top \\ \implies 0 &= 4W(H - H^*)W - \vec{\lambda}\vec{y}^\top - \vec{y}\vec{\lambda}^\top\end{aligned}$$

Post-multiplying this relationship by $\vec{s}$ shows:

$$\vec{0} = 4(\vec{y} - WH^*\vec{y}) - \vec{\lambda}(\vec{y} \cdot \vec{s}) - \vec{y}(\vec{\lambda} \cdot \vec{s})$$

Now, take the dot product with $\vec{s}$:

$$0 = 4(\vec{y} \cdot \vec{s}) - 4(\vec{y}^\top H^*\vec{y}) - 2(\vec{y} \cdot \vec{s})(\vec{\lambda} \cdot \vec{s})$$

This shows:

$$\vec{\lambda} \cdot \vec{s} = 2\rho\vec{y}^\top(\vec{s} - H^*\vec{y}), \text{ for } \rho \equiv 1/\vec{y}\cdot\vec{s}$$

Now, we substitute this into our vector equality:

$$\begin{aligned}\vec{0} &= 4(\vec{y} - WH^*\vec{y}) - \vec{\lambda}(\vec{y} \cdot \vec{s}) - \vec{y}(\vec{\lambda} \cdot \vec{s}) \text{ from before} \\ &= 4(\vec{y} - WH^*\vec{y}) - \vec{\lambda}(\vec{y} \cdot \vec{s}) - \vec{y}[2\rho\vec{y}^\top(\vec{s} - H^*\vec{y})] \text{ from our simplification} \\ \implies \vec{\lambda} &= 4\rho(\vec{y} - WH^*\vec{y}) - 2\rho^2\vec{y}^\top(\vec{s} - H^*\vec{y})\vec{y}\end{aligned}$$

*Special thanks to Tao Du for debugging several parts of this derivation.

Post-multiplying by $\vec{y}^\top$ shows:

$$\vec{\lambda}\vec{y}^\top = 4\rho(\vec{y} - WH^*\vec{y})\vec{y}^\top - 2\rho^2\vec{y}^\top(\vec{s} - H^*\vec{y})\vec{y}\vec{y}^\top$$

Taking the transpose,

$$\vec{y}\vec{\lambda}^\top = 4\rho\vec{y}(\vec{y}^\top - \vec{y}^\top H^*W) - 2\rho^2\vec{y}^\top(\vec{s} - H^*\vec{y})\vec{y}\vec{y}^\top$$

Combining these results and dividing by four shows:

$$\frac{1}{4}(\vec{\lambda}\vec{y}^\top + \vec{y}\vec{\lambda}^\top) = \rho(2\vec{y}\vec{y}^\top - WH^*\vec{y}\vec{y}^\top - \vec{y}\vec{y}^\top H^*W) - \rho^2\vec{y}^\top(\vec{s} - H^*\vec{y})\vec{y}\vec{y}^\top$$

Now, we will pre- and post-multiply by W^{-1}. Since $W\vec{s} = \vec{y}$, we can equivalently write $\vec{s} = W^{-1}\vec{y}$; furthermore, by symmetry of W we then know $\vec{y}^\top W^{-1} = \vec{s}^\top$. Applying these identities to the expression above shows:

$$\begin{aligned}
\frac{1}{4}W^{-1}(\vec{\lambda}\vec{y}^\top + \vec{y}\vec{\lambda}^\top)W^{-1} &= 2\rho\vec{s}\vec{s}^\top - \rho H^*\vec{y}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^* - \rho^2(\vec{y}^\top\vec{s})\vec{s}\vec{s}^\top + \rho^2(\vec{y}^\top H^*\vec{y})\vec{s}\vec{s}^\top \\
&= 2\rho\vec{s}\vec{s}^\top - \rho H^*\vec{y}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^* - \rho\vec{s}\vec{s}^\top + \vec{s}\rho^2(\vec{y}^\top H^*\vec{y})\vec{s}^\top \text{ by definition of } \rho \\
&= \rho\vec{s}\vec{s}^\top - \rho H^*\vec{y}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^* + \vec{s}\rho^2(\vec{y}^\top H^*\vec{y})\vec{s}^\top
\end{aligned}$$

Finally, we can conclude our derivation of the BFGS step as follows:

$$\begin{aligned}
0 &= 4W(H - H^*)W - \vec{\lambda}\vec{y}^\top - \vec{y}\vec{\lambda}^\top \text{ from before} \\
\implies H &= \frac{1}{4}W^{-1}(\vec{\lambda}\vec{y}^\top + \vec{y}\vec{\lambda}^\top)W^{-1} + H^* \\
&= \rho\vec{s}\vec{s}^\top - \rho H^*\vec{y}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^* + \vec{s}\rho^2(\vec{y}^\top H^*\vec{y})\vec{s}^\top + H^* \text{ from the last paragraph} \\
&= H^*(I - \rho\vec{y}\vec{s}^\top) + \rho\vec{s}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^* + (\rho\vec{s}\vec{y}^\top)H^*(\rho\vec{y}\vec{s}^\top) \\
&= H^*(I - \rho\vec{y}\vec{s}^\top) + \rho\vec{s}\vec{s}^\top - \rho\vec{s}\vec{y}^\top H^*(I - \rho\vec{y}\vec{s}^\top) \\
&= \rho\vec{s}\vec{s}^\top + (I - \rho\vec{s}\vec{y}^\top)H^*(I - \rho\vec{y}\vec{s}^\top)
\end{aligned}$$

This final expression is exactly the BFGS step introduced in the chapter.

CHAPTER 9

Constrained Optimization

CONTENTS

9.1 Motivation 178
9.2 Theory of Constrained Optimization 181
9.2.1 Optimality 181
9.2.2 KKT Conditions 181
9.3 Optimization Algorithms 184
9.3.1 Sequential Quadratic Programming (SQP) 185
9.3.1.1 Equality constraints 185
9.3.1.2 Inequality Constraints 185
9.3.2 Barrier Methods 186
9.4 Convex Programming 186
9.4.1 Linear Programming 188
9.4.2 Second-Order Cone Programming 189
9.4.3 Semidefinite Programming 191
9.4.4 Integer Programs and Relaxations 193

We continue our consideration of optimization problems by studying the *constrained* case. These problems take the following general form:

$$\begin{aligned} &\text{minimize } f(\vec{x}) \\ &\text{such that } g(\vec{x}) = \vec{0} \\ &\qquad\qquad h(\vec{x}) \geq \vec{0} \end{aligned}$$

Here, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$; we generally call f the *objective function* and the expressions $g(\vec{x}) = \vec{0}, h(\vec{x}) \geq \vec{0}$ the *constraints*. This form is extremely generic, so algorithms for solving such problems in the absence of additional assumptions on f, g, or h can be difficult to formulate and are subject to degeneracies such as local minima and lack of convergence. In fact, this general problem encodes other problems we already have considered. If we take $f(\vec{x}) = h(\vec{x}) \equiv 0$, then this constrained optimization becomes root-finding on g (Chapter 7), while if we take $g(\vec{x}) = h(\vec{x}) \equiv \vec{0}$ then it reduces to unconstrained optimization on f (Chapter 8).

Despite this bleak outlook, optimization methods handling the general constrained case can be valuable when f, g, and h do not have useful structure. Furthermore, when f is heuristic anyway, simply finding a feasible $\vec{x}$ for which $f(\vec{x}) < f(\vec{x}_0)$ for an initial guess $\vec{x}_0$ can be valuable. One application in this domain would be an economic system in which f measures costs; we wish to minimize costs, but if $\vec{x}_0$ represents the current configuration, *any* $\vec{x}$ decreasing f is a valuable output.

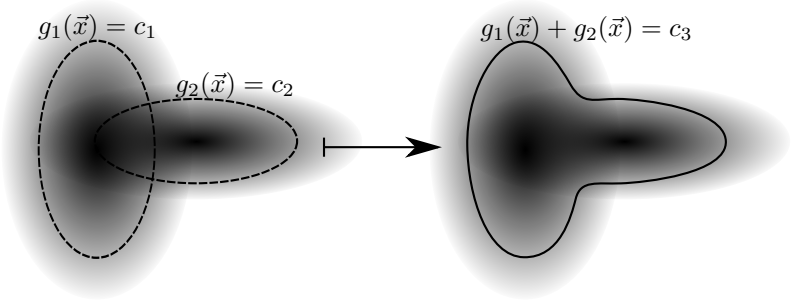


FIGURE 9.1 "Blobby" shapes are constructed as level sets of a linear combination of functions.

9.1 MOTIVATION

Constrained optimization problems appear in nearly any area of applied math, engineering, and computer science. We already listed many applications of constrained optimization when we discussed eigenvectors and eigenvalues in Chapter 5, since this problem for symmetric matrices can be posed as finding critical points of $\vec{x}^\top A\vec{x}$ subject to $\|\vec{x}\|_2 = 1$. The particular case of eigenvalue computation admits special algorithms that make it a simpler problem. Here, however, we list other optimization problems that do not enjoy the unique structure of eigenvalue problems:

Example 9.1 (Geometric projection)**.** Many shapes S in $\mathbb{R}^n$ can be written *implicitly* in the form $g(\vec{x}) = 0$ for some g. For example, the unit sphere results from taking $g(\vec{x}) \equiv \|\vec{x}\|_2^2 - 1$, while a cube can be constructed by taking $g(\vec{x}) = \|\vec{x}\|_1 - 1$. Some 3D modeling environments allow users to specify "blobby" objects, as in Figure 9.1, as zero level sets of $g(\vec{x})$ given by

$$g(\vec{x}) \equiv c + \sum_i a_i e^{-b_i \|\vec{x} - \vec{x}_i\|_2^2}.$$

Suppose we are given a point $\vec{y} \in \mathbb{R}^3$ and wish to find the closest point on S to $\vec{y}$. This problem is solved by using the following constrained minimization:

$$\begin{aligned} &\text{minimize}_{\vec{x}} \ \|\vec{x} - \vec{y}\|_2 \\ &\text{such that } g(\vec{x}) = 0 \end{aligned}$$

Example 9.2 (Manufacturing)**.** Suppose you have m different materials; you have s_i units of each material i in stock. You can manufacture k different products; product j gives you profit p_j and uses c_{ij} of material i to make. To maximize profits, you can solve the following optimization for the amount x_j you should manufacture of each item j:

$$\begin{aligned} &\text{maximize}_{\vec{x}} \ \sum_{j=1}^{k} p_j x_j \\ &\text{such that } x_j \geq 0 \, \forall j \in \{1, \ldots, k\} \\ &\sum_{j=1}^{k} c_{ij} x_j \leq s_i \, \forall i \in \{1, \ldots, m\} \end{aligned}$$

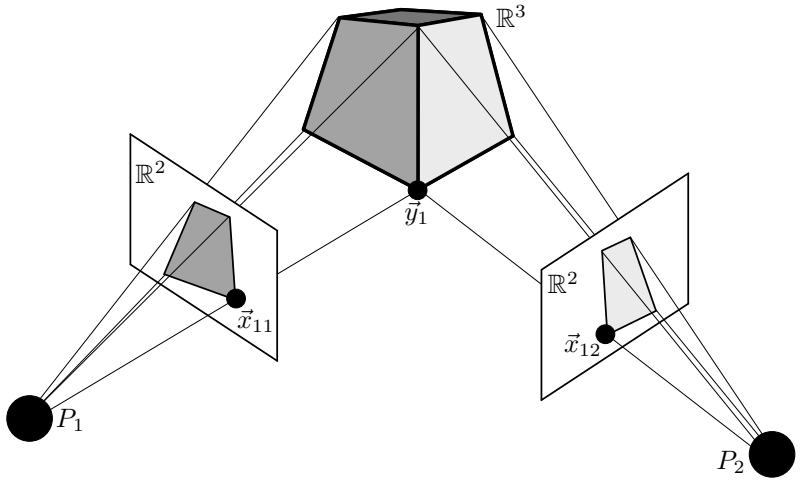


FIGURE 9.2 Notation for bundle adjustment with two images. Given corresponding points $\vec{x}_{ij}$ marked on images, bundle adjustment simultaneously optimizes for camera parameters encoded in P_i and three-dimensional positions $\vec{y}_j$.

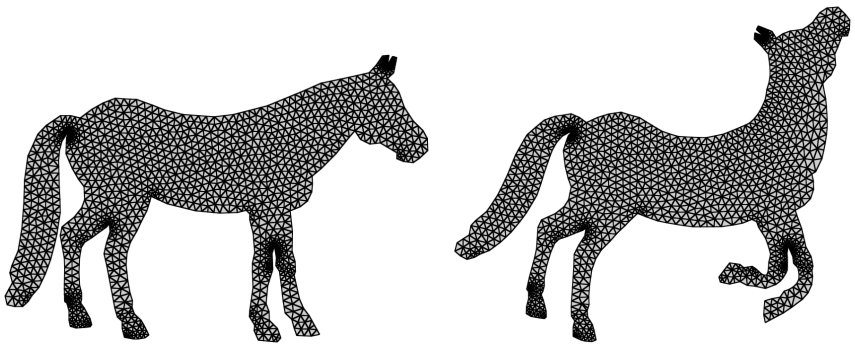
The first constraint ensures that you do not make negative numbers of any product, and the second ensures that you do not use more than your stock of each material. This optimization is an example of a *linear program*, because the objective and constraints are all linear functions.

Example 9.3 (Nonnegative least-squares)**.** We already have seen numerous examples of least-squares problems, but sometimes negative values in the solution vector might not make sense. For example, in computer graphics, an animated model might be expressed as a deforming bone structure plus a meshed "skin;" for each point on the skin a list of weights can be computed to approximate the influence of the positions of the bone joints on the position of the skin vertices [62]. Such weights should be constrained to be nonnegative to avoid degenerate behavior while the surface deforms. In such a case, we can solve the "nonnegative least-squares" problem:

$$\begin{aligned} &\text{minimize}_{\vec{x}} \ \|A\vec{x} - \vec{b}\|_2 \\ &\text{such that } x_i \geq 0 \ \forall i \end{aligned}$$

Machine learning methods leverage the *sparsity* of nonnegative least squares solutions, which often satisfy $x_i = 0$ for many indices i [108].

Example 9.4 (Bundle adjustment)**.** In computer vision, suppose we take pictures of an object from several angles. A natural task is to reconstruct the three-dimensional shape of the object. To do so, we might mark a corresponding set of points on each image; in particular, we can take $\vec{x}_{ij} \in \mathbb{R}^2$ to be the position of feature point j on image i, as illustrated in Figure 9.2. In reality, each feature point has a position $\vec{y}_j \in \mathbb{R}^3$ in space, which we would like to compute. Additionally, we must find the positions of the cameras themselves, which we can represent as unknown projection matrices P_i. This problem,



(a) Original (b) Deformed

FIGURE 9.3 As-rigid-as-possible (ARAP) optimization generates the deformed mesh on the right from the original mesh on the left given target positions for a few points on the head, feet, and torso.

known as *bundle adjustment*, can be posed as an optimization:

$$\text{minimize}_{\vec{y}_j, P_i} \sum_{ij} \|P_i \vec{y}_j - \vec{x}_{ij}\|_2^2$$
$$\text{such that } P_i \text{ is orthogonal } \forall i.$$

The orthogonality constraint ensures that the camera transformations are reasonable.

Example 9.5 (As-rigid-as-possible deformation)**.** The "as-rigid-as-possible" (ARAP) modeling technique is used in computer graphics to deform two- and three-dimensional shapes in real time for modeling and animation software [110]. In the planar setting, suppose we are given a two-dimensional triangle mesh, as in Figure 9.3(a). This mesh consists of a collection of vertices V connected into triangles by edges $E \subseteq V \times V$; we will assume each vertex $v \in V$ is associated with a position $\vec{x}_v \in \mathbb{R}^2$. Furthermore, assume the user manually moves a subset of vertices $V_0 \subset V$ to target positions $\vec{y}_v \in \mathbb{R}^2$ for $v \in V_0$ to specify a potential deformation of the shape. The goal of ARAP is to deform the remainder $V \backslash V_0$ of the mesh vertices elastically, as in Figure 9.3(b), yielding a set of new positions $\vec{y}_v \in \mathbb{R}^2$ for each $v \in V$ with $\vec{y}_v$ fixed by the user when $v \in V_0$.

The least-distorting deformation of the mesh is a *rigid* motion, meaning it rotates and translates but does not stretch or shear. In this case, there exists an *orthogonal* matrix $R \in \mathbb{R}^{2\times 2}$ so that the deformation satisfies $\vec{y}_v - \vec{y}_w = R(\vec{x}_v - \vec{x}_w)$ for any edge $(v, w) \in E$. But, if the user wishes to stretch or bend part of the shape, there might not exist a single R rotating the entire mesh to satisfy the position constraints in V_0.

To loosen the single-rotation assumption, ARAP asks that a deformation is *approximately* or *locally* rigid. Specifically, no single vertex on the mesh should experience more than a little stretch or shear, so in a neighborhood of each vertex $v \in V$ there should exist an orthogonal matrix R_v satisfying $\vec{y}_v - \vec{y}_w \approx R_v(\vec{x}_v - \vec{x}_w)$ for any $(v, w) \in E$. Once again applying least-squares, we define the as-rigid-as-possible deformation of the mesh to be the one mapping $\vec{x}_v \mapsto \vec{y}_v$ for all $v \in V$ by solving the following optimization

problem:

$$\begin{aligned}\text{minimize}_{R_v,\vec{y}_v} &\sum_{v\in V}\sum_{(v,w)\in E}\|R_v(\vec{x}_v-\vec{x}_w)-(\vec{y}_v-\vec{y}_w)\|_2^2\\ \text{such that } &R_v^\top R_v = I_{2\times 2}\ \forall v\in V\\ &\vec{y}_v \text{ fixed } \forall v\in V_0\end{aligned}$$

We will suggest one way to solve this optimization problem in Example 11.5.

9.2 THEORY OF CONSTRAINED OPTIMIZATION

In our discussion, we will assume that f, g, and h are differentiable. Some methods exist that only make weak continuity or Lipschitz assumptions, but these techniques are quite specialized and require advanced analytical consideration.

9.2.1 Optimality

Although we have not yet developed algorithms for general constrained optimization, we implicitly have made use of the *theory* of such problems while considering many of the problems listed at the beginning of Chapter 8. Specifically, recall the method of Lagrange multipliers, introduced in Theorem 0.1. In this technique, critical points of $f(\vec{x})$ subject to $g(\vec{x})=\vec{0}$ are given by critical points of the *unconstrained* Lagrange multiplier function

$$\Lambda(\vec{x},\vec{\lambda}) \equiv f(\vec{x}) - \vec{\lambda}\cdot\vec{g}(\vec{x})$$

with respect to both $\vec{\lambda}$ and $\vec{x}$ simultaneously. This theorem allowed us to provide variational interpretations of eigenvalue problems; more generally, it gives an alternative criterion for $\vec{x}$ to be a critical point of an *equality-constrained* optimization.

As we saw in Chapter 7, simply finding an $\vec{x}$ satisfying the constraint $g(\vec{x})=\vec{0}$, however, can be a considerable challenge independently from minimizing $f(\vec{x})$. We can separate these issues by making a few definitions:

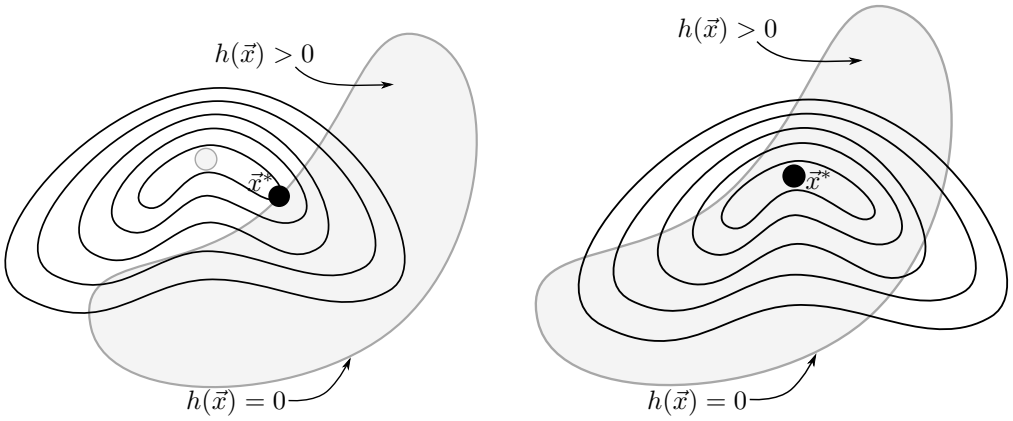
Definition 9.1 (Feasible point and feasible set)**.** A *feasible point* of a constrained optimization problem is any point $\vec{x}$ satisfying $g(\vec{x})=\vec{0}$ and $h(\vec{x})\geq\vec{0}$. The *feasible set* is the set of all points $\vec{x}$ satisfying these constraints.

Definition 9.2 (Critical point of constrained optimization)**.** A critical point of a constrained optimization is one satisfying the constraints that also is a local maximum, minimum, or saddle point of f within the feasible set.

9.2.2 KKT Conditions

Constrained optimizations are difficult because they simultaneously solve root-finding problems (the $g(\vec{x})=\vec{0}$ constraint), satisfiability problems (the $h(\vec{x})\geq\vec{0}$ constraint), and minimization (on the function f). As stated in Theorem 0.1, Lagrange multipliers allow us to turn equality-constrained minimization problems into root-finding problems on Λ. To push our differential techniques to complete generality, however, we must find a way to add inequality constraints $h(\vec{x})\geq\vec{0}$ to the Lagrange multiplier system.

Suppose we have found a local minimum subject to the constraints, denoted $\vec{x}^*$. For each inequality constraint $h_i(\vec{x}^*)\geq 0$, we have two options:



Active constraint
$h(\vec{x}^*) = 0$

Inactive constraint
$h(\vec{x}^*) > 0$

FIGURE 9.4 Active and inactive constraints $h(\vec{x}) \geq 0$ for minimizing a function whose level sets are shown in black; the region $h(\vec{x}) \geq 0$ is shown in gray. When the $h(\vec{x}) \geq 0$ constraint is active, the optimal point $\vec{x}^*$ is on the border of the feasible domain and would move if the constraint were removed. When the constraint is inactive, $\vec{x}^*$ is in the *interior* of the feasible set, so the constraint $h(\vec{x}) \geq 0$ has no effect on the position of the $\vec{x}^*$ locally.

- $h_i(\vec{x}^*) = 0$: Such a constraint is *active*, likely indicating that if the constraint were removed $\vec{x}^*$ would no longer be optimal.
- $h_i(\vec{x}^*) > 0$: Such a constraint is *inactive*, meaning in a neighborhood of $\vec{x}^*$ if we had removed this constraint we still would have reached the same minimum.

These two cases are illustrated in Figure 9.4. While this classification will prove valuable, we do not know *a priori* which constraints will be active or inactive at $\vec{x}^*$ until we solve the optimization problem and find $\vec{x}^*$.

If all of our constraints were active, then we could change the constraint $h(\vec{x}) \geq \vec{0}$ to an equality without affecting the outcome of the optimization. If this were the case, then by the equality-constrained Lagrange multiplier conditions we could study the following Lagrange multiplier expression:

$$\Lambda(\vec{x}, \vec{\lambda}, \vec{\mu}) \equiv f(\vec{x}) - \vec{\lambda} \cdot g(\vec{x}) - \vec{\mu} \cdot h(x)$$

We no longer can say that $\vec{x}^*$ is a critical point of Λ, however, because inactive constraints would remove terms above. Ignoring this (important!) issue for the time being, we could proceed blindly and ask for critical points of this new Λ with respect to $\vec{x}$, which satisfy the following:

$$\vec{0} = \nabla f(\vec{x}) - \sum_i \lambda_i \nabla g_i(\vec{x}) - \sum_j \mu_j \nabla h_j(\vec{x})$$

Here we have separated out the individual components of g and h and treated them as scalar functions to avoid complex notation.

A clever trick can extend this optimality condition to inequality constraints. If we *define*

$\mu_j \equiv 0$ whenever h_j is inactive, then the irrelevant terms are removed from the optimality conditions. In other words, we can *add* a constraint on the Lagrange multiplier above:

$$\mu_j h_j(\vec{x}) = 0.$$

With this constraint in place, we know that at least one of μ_j and $h_j(\vec{x})$ must be zero; when the constraint $h_j(\vec{x}) \geq 0$ is inactive, then μ_j must equal zero to compensate. In other words, our first-order optimality condition still holds at critical points of the inequality-constrained problem—after adding this extra constraint.

So far, our construction has not distinguished between the constraint $h_j(\vec{x}) \geq 0$ and the constraint $h_j(\vec{x}) \leq 0$. If the constraint is inactive, it could have been dropped without affecting the outcome of the optimization locally, so we consider the case when the constraint is active. Intuitively,* in this case we expect there to be a way to decrease f by violating the constraint. Locally, the direction in which f decreases is $-\nabla f(\vec{x}^*)$ and the direction in which h_j decreases is $-\nabla h_j(\vec{x}^*)$. Thus, starting at $\vec{x}^*$ we can decrease f even more by violating the constraint $h_j(\vec{x}) \geq 0$ when $\nabla f(\vec{x}^*) \cdot \nabla h_j(\vec{x}^*) \geq 0$.

Products of gradients of f and h_j are difficult to manipulate. At $\vec{x}^*$, however, our first-order optimality condition tells us:

$$\nabla f(\vec{x}^*) = \sum_i \lambda_i^* \nabla g_i(\vec{x}^*) + \sum_{j \text{ active}} \mu_j^* \nabla h_j(\vec{x}^*)$$

The inactive μ_j values are zero and can be removed. We can remove the $g(\vec{x}) = 0$ constraints by adding inequality constraints $g(\vec{x}) \geq \vec{0}$ and $g(\vec{x}) \leq \vec{0}$ to h; this is a mathematical convenience for writing a proof rather than a numerically-wise maneuver. Then, taking dot products with ∇h_k for any fixed k shows:

$$\sum_{j \text{ active}} \mu_j^* \nabla h_j(\vec{x}^*) \cdot \nabla h_k(\vec{x}^*) = \nabla f(\vec{x}^*) \cdot \nabla h_k(\vec{x}^*) \geq 0$$

Vectorizing this expression shows $Dh(\vec{x}^*)Dh(\vec{x}^*)^\top \vec{\mu}^* \geq \vec{0}$. Since $Dh(\vec{x}^*)Dh(x^*)^\top$ is positive semidefinite, this implies $\vec{\mu}^* \geq \vec{0}$. Thus, the $\nabla f(\vec{x}^*) \cdot \nabla h_j(\vec{x}^*) \geq 0$ observation is equivalent to the much easier condition $\mu_j \geq 0$.

Our observations can be combined and formalized to prove a first-order optimality condition for inequality-constrained minimization problems:

Theorem 9.1 (Karush-Kuhn-Tucker (KKT) conditions)**.** The vector $\vec{x}^* \in \mathbb{R}^n$ is a critical point for minimizing f subject to $g(\vec{x}) = \vec{0}$ and $h(\vec{x}) \geq \vec{0}$ when there exists $\vec{\lambda} \in \mathbb{R}^m$ and $\vec{\mu} \in \mathbb{R}^p$ such that:

- $\vec{0} = \nabla f(\vec{x}^*) - \sum_i \lambda_i \nabla g_i(\vec{x}^*) - \sum_j \mu_j \nabla h_j(\vec{x}^*)$ ("stationarity")
- $g(\vec{x}^*) = \vec{0}$ and $h(\vec{x}^*) \geq \vec{0}$ ("primal feasibility")
- $\mu_j h_j(\vec{x}^*) = 0$ for all j ("complementary slackness")
- $\mu_j \geq 0$ for all j ("dual feasibility")

When h is removed this theorem reduces to the Lagrange multiplier criterion.

*You should not consider our discussion a formal proof, since we are not considering many boundary cases.

Example 9.6 (KKT conditions)**.** Suppose we wish to solve the following optimization (proposed by R. Israel, UBC Math 340, Fall 2006):

$$\begin{aligned} &\text{maximize } xy \\ &\text{such that } x + y^2 \leq 2 \\ &\qquad x, y \geq 0 \end{aligned}$$

In this case we will have no λ's and three μ's. We take $f(x, y) = -xy$, $h_1(x, y) \equiv 2-x-y^2$, $h_2(x, y) = x$, and $h_3(x, y) = y$. The KKT conditions are:

$$\begin{aligned} \text{Stationarity: } & 0 = -y + \mu_1 - \mu_2 \\ & 0 = -x + 2\mu_1 y - \mu_3 \\ \text{Primal feasibility: } & x + y^2 \leq 2 \\ & x, y \geq 0 \\ \text{Complementary slackness: } & \mu_1(2 - x - y^2) = 0 \\ & \mu_2 x = 0 \\ & \mu_3 y = 0 \\ \text{Dual feasibility: } & \mu_1, \mu_2, \mu_3 \geq 0 \end{aligned}$$

Example 9.7 (Linear programming)**.** Consider the optimization:

$$\begin{aligned} &\text{minimize}_{\vec{x}} \ \vec{b} \cdot \vec{x} \\ &\text{such that } A\vec{x} \geq \vec{c} \end{aligned}$$

Notice Example 9.2 can be written this way. The KKT conditions for this problem are:

$$\begin{aligned} \text{Stationarity: } & A^\top \vec{\mu} = \vec{b} \\ \text{Primal feasibility: } & A\vec{x} \geq \vec{c} \\ \text{Complementary slackness: } & \mu_i(\vec{a}_i \cdot \vec{x} - c_i) = 0 \ \forall i, \text{ where } \vec{a}_i^\top \text{ is row } i \text{ of } A \\ \text{Dual feasibility: } & \vec{\mu} \geq \vec{0} \end{aligned}$$

As with Lagrange multipliers, we cannot assume that any $\vec{x}^*$ satisfying the KKT conditions automatically minimizes f subject to the constraints, even locally. One way to check for local optimality is to examine the Hessian of f restricted to the subspace of $\mathbb{R}^n$ in which $\vec{x}$ can move without violating the constraints; if this "reduced" Hessian is positive definite then the optimization has reached a local minimum.

9.3 OPTIMIZATION ALGORITHMS

A careful consideration of algorithms for constrained optimization is out of the scope of our discussion; thankfully many stable implementations of these techniques exist and much can be accomplished as a "client" of this software rather than rewriting it from scratch. Even so, it is useful to sketch some common approaches to gain some intuition for how these libraries work.

9.3.1 Sequential Quadratic Programming (SQP)

Similar to BFGS and other methods we considered in Chapter 8, one typical strategy for constrained optimization is to approximate f, g, and h with simpler functions, solve the approximate optimization, adjust the approximation based on the latest function evaluation, and repeat.

Suppose we have a guess $\vec{x}_k$ of the solution to the constrained optimization problem. We could apply a second-order Taylor expansion to f and first-order approximation to g and h to define a next iterate as the following:

$$\begin{aligned}\vec{x}_{k+1} \equiv \vec{x}_k + \arg\min_{\vec{d}} \; & \left[\frac{1}{2}\vec{d}^\top H_f(\vec{x}_k)\vec{d} + \nabla f(\vec{x}_k)\cdot\vec{d} + f(\vec{x}_k)\right] \\ \text{such that } & g_i(\vec{x}_k) + \nabla g_i(\vec{x}_k)\cdot\vec{d} = 0 \\ & h_i(\vec{x}_k) + \nabla h_i(\vec{x}_k)\cdot\vec{d} \geq 0\end{aligned}$$

The optimization to find $\vec{d}$ has a quadratic objective with linear constraints, which can be carried out using one of many strategies; it is known as a *quadratic program*. This Taylor approximation, however, only works in a neighborhood of the optimal point; when a good initial guess $\vec{x}_0$ is unavailable, these strategies may fail.

9.3.1.1 Equality constraints

When the only constraints are equalities and h is removed, the quadratic program for $\vec{d}$ has Lagrange multiplier optimality conditions derived as follows:

$$\begin{aligned}\Lambda(\vec{d},\vec{\lambda}) &\equiv \frac{1}{2}\vec{d}^\top H_f(\vec{x}_k)\vec{d} + \nabla f(\vec{x}_k)\cdot\vec{d} + f(\vec{x}_k) + \vec{\lambda}^\top(g(\vec{x}_k) + Dg(\vec{x}_k)\vec{d}) \\ \implies \vec{0} = \nabla_{\vec{d}}\Lambda &= H_f(\vec{x}_k)\vec{d} + \nabla f(\vec{x}_k) + [Dg(\vec{x}_k)]^\top\vec{\lambda}\end{aligned}$$

Combining this expression with the earlier equality condition yields a symmetric linear system for $\vec{d}$ and $\vec{\lambda}$:

$$\begin{pmatrix} H_f(\vec{x}_k) & [Dg(\vec{x}_k)]^\top \\ Dg(\vec{x}_k) & 0 \end{pmatrix}\begin{pmatrix}\vec{d}\\ \vec{\lambda}\end{pmatrix} = \begin{pmatrix} -\nabla f(\vec{x}_k) \\ -g(\vec{x}_k)\end{pmatrix}$$

Thus, each iteration of sequential quadratic programming in the presence of only equality constraints can be accomplished by solving this linear system at each iteration to get $\vec{x}_{k+1} \equiv \vec{x}_k + \vec{d}$. This linear system above is *not* positive definite, so on a large scale it can be difficult to solve.

Extensions operate like BFGS for unconstrained optimization by approximating the Hessian H_f. Stability also can be improved by limiting the distance that $\vec{x}$ can move during any single iteration.

9.3.1.2 Inequality Constraints

Specialized algorithms exist for solving quadratic programs rather than general nonlinear programs, and these can be used to generate steps of SQP. One notable strategy is to keep an "active set" of constraints that are active at the minimum with respect to $\vec{d}$. Then, the equality-constrained methods above can be applied by ignoring inactive constraints. Iterations of active-set optimization update the active set by adding violated constraints and removing those inequality constraints h_j for which $\nabla f \cdot \nabla h_j \leq 0$ as in §9.2.2.

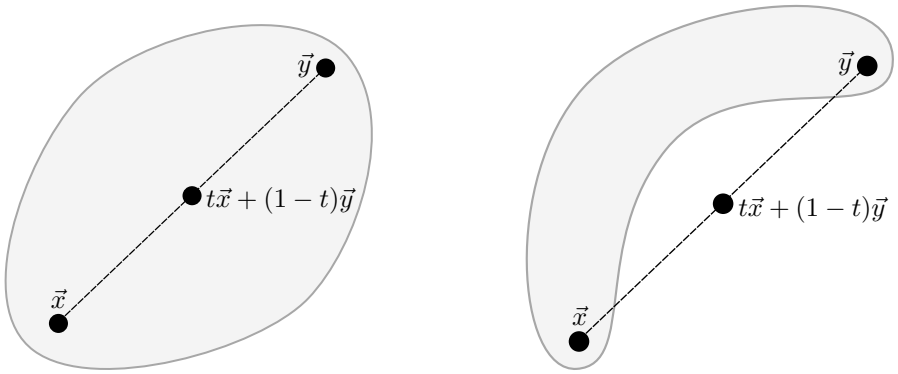


FIGURE 9.5 Convex and nonconvex shapes on the plane.

9.3.2 Barrier Methods

Another option for minimization in the presence of constraints is to change the constraints to energy terms. For example, in the equality constrained case we could minimize an "augmented" objective as follows:

$$f_\rho(\vec{x}) = f(\vec{x}) + \rho\|g(\vec{x})\|_2^2$$

Taking $\rho \to \infty$ will force $g(\vec{x})$ to be as small as possible, so eventually we will reach $g(\vec{x}) \approx \vec{0}$. Thus, the *barrier method* of constrained optimization applies iterative *unconstrained* optimization techniques to f_ρ and checks how well the constraints are satisfied; if they are not within a given tolerance, ρ is increased and the optimization continues using the previous iterate as a starting point.

Barrier methods are simple to implement and use, but they can exhibit some pernicious failure modes. In particular, as ρ increases, the influence of f on the objective function diminishes and the Hessian of f_ρ becomes more and more poorly-conditioned.

Barrier methods can be applied to inequality constraints as well. In this case, we must ensure that $h_i(\vec{x}) \geq 0$ for all i. Typical choices of barrier functions might include $1/h_i(\vec{x})$—the "inverse barrier"—and $-\log h_i(\vec{x})$—the "logarithmic barrier."

9.4 CONVEX PROGRAMMING

The methods we have described for constrained optimization come with few guarantees on the quality of the output. Certainly they are unable to obtain global minima without a good initial guess $\vec{x}_0$, and in certain cases, e.g. when Hessians near $\vec{x}^*$ is not positive definite, they may not converge at all.

There is one notable exception to this rule, which appears in any number of optimizations: *convex programming*. The idea here is that when f is a convex function and the feasible set itself is convex, then the optimization possesses a unique minimum. We already have defined a convex function, but need to understand what it means for a set of constraints to be convex:

Definition 9.3 (Convex set)**.** A set $S \subseteq \mathbb{R}^n$ is *convex* if for any $\vec{x}, \vec{y} \in S$, the point $t\vec{x} + (1-t)\vec{y}$ is also in S for any $t \in [0, 1]$.

As shown in Figure 9.5, intuitively a set is convex if its boundary shape does not bend inward and outward.

Example 9.8 (Circles)**.** The disc $\{\vec{x} \in \mathbb{R}^n : \|\vec{x}\|_2 \leq 1\}$ is convex, while the unit circle $\{\vec{x} \in \mathbb{R}^n : \|\vec{x}\|_2 = 1\}$ is not.

A nearly identical proof to that of Proposition 8.1 shows:

A convex function cannot have suboptimal local minima even when it is restricted to a convex domain.

In particular, if the convex objective function has two local minima, then the line of points between those minima must yield objective values less than or equal to those on the endpoints; by Definition 9.3 this entire line is feasible, completing the proof.

Strong convergence guarantees are available for convex optimization methods that guarantee finding a *global* minimum so long as f is convex and the constraints on g and h make a convex feasible set. Thus, a valuable exercise for nearly any optimization problem is to check if it is convex, since such an observation can increase confidence in the solution quality and the chances of success by a large factor.

A new field called *disciplined convex programming* attempts to chain together rules about convexity to generate convex optimizations, allowing the end user to combine convex energy terms and constraints so long as they satisfy criteria making the final optimization convex. Useful statements about convexity in this domain include the following:

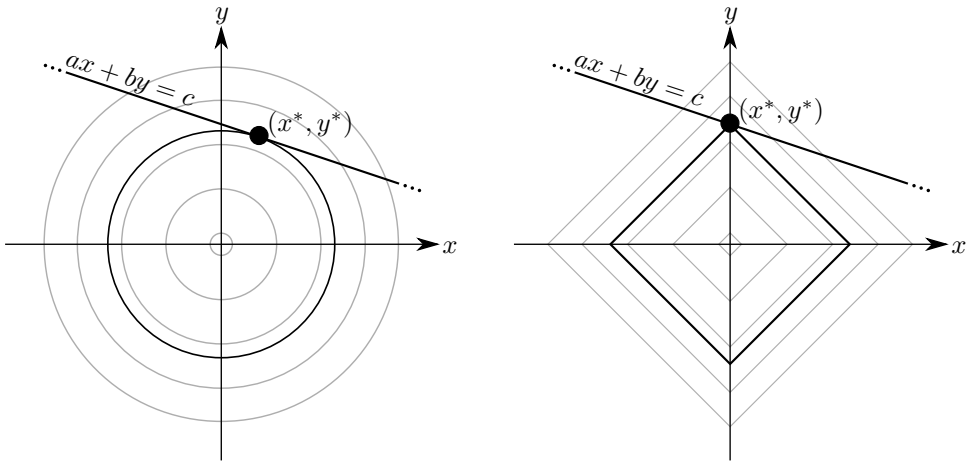
- The intersection of convex sets is convex; thus, adding multiple convex constraints is an allowable operation.
- The sum of convex functions is convex.
- If f and g are convex, so is $h(\vec{x}) \equiv \max\{f(\vec{x}), g(\vec{x})\}$.
- If f is a convex function, the set $\{\vec{x} : f(\vec{x}) \leq c\}$ is convex.

Tools such as the `CVX` library help separate implementation of assorted convex objectives from their minimization [47, 48].

Example 9.9 (Convex programming)**.**

- The nonnegative least squares problem in Example 9.3 is convex because $\|A\vec{x} - \vec{b}\|_2$ is a convex function of $\vec{x}$ and the set $\vec{x} \geq \vec{0}$ is convex.
- Linear programming, as introduced in in Example 9.7, is convex because it has a linear objective and linear constraints.
- We can include $\|\vec{x}\|_1$ in a convex optimization objective by introducing a variable $\vec{y}$. To do so, we add constraints $y_i \geq x_i$ and $y_i \geq -x_i$ for each i and an objective $\sum_i y_i$. This sum has terms that are at least as large as $|x_i|$ and that the energy and constraints are convex. At the minimum we must have $y_i = |x_i|$ since we have constrained $y_i \geq |x_i|$ and we wish to minimize the energy. "Disciplined" convex libraries can do such operations behind the scenes without revealing such substitutions to the end user.

Convex programming has much in common with aspects of computer science theory involving *reductions* of algorithmic problems to one another. Rather than verifying NP-completeness, however, in this context we wish to use a generic "solver" to optimize given



(a) $p = 2$ (b) $p = 1$

FIGURE 9.6 On the (x, y) plane, the optimization minimizing $\|(x, y)\|_p$ subject to $ax + by = c$ has considerably different output depending on whether we choose $p = 2$ or $p = 1$. Level sets $\{(x, y) : \|(x, y)\|_p = c\}$ are shown in gray.

convex objective, just like we reduced assorted computational problems to a linear solve in Chapter 3. There is a formidable pantheon of industrial-scale convex programming tools that can handle different classes of problems with varying levels of efficiency and generality; below we detail some of the most common classes with examples of how they might be applied. See [13, 79] for approachable extended discussions of convex programming and related topics.

9.4.1 Linear Programming

A particularly well-studied example of convex optimization is *linear programing*, originally introduced in Example 9.7. Exercise 9.4 will walk through the derivation of some unique properties making linear programs attractive both theoretically and from an algorithmic design standpoint.

The famous *simplex algorithm*, which can be considered an active set method as in §9.3.1.2, solves for the resulting $\vec{x}^*$ using a linear system, and checks if the active set must be updated. No Taylor approximations are needed because the objective and feasible set are given by linear expressions. *Interior point* linear programming strategies such as the barrier method also are successful for these problems. For this reason, linear programs can be solved on a *huge* scale—up to millions or billions of variables!—and often appear in problems like scheduling or pricing.

One popular application of linear programming inspired by Example 9.9 provides an alternative to using pseudoinverse for underdetermined linear systems (§6.2.1). When a matrix A is underdetermined, there are many vectors $\vec{x}$ that could satisfy $A\vec{x} = \vec{b}$ for a given vector $\vec{b}$. In this case, the pseudoinverse A^+ applied to $\vec{b}$ effectively solves the following optimization problem:

$$\text{Pseudoinverse} \left\{ \begin{array}{ll} \text{minimize}_{\vec{x}} & \|\vec{x}\|_2 \\ \text{such that} & A\vec{x} = \vec{b} \end{array} \right.$$

Using linear programs, we can solve a slightly different system:

$$L_1 \text{ minimization} \begin{cases} \text{minimize}_{\vec{x}} & \|\vec{x}\|_1 \\ \text{such that} & A\vec{x} = \vec{b} \end{cases}$$

All we have done here is replace the norm $\|\cdot\|_2$ with a different norm $\|\cdot\|_1$.

Why does this one-character change make a significant difference in the output $\vec{x}$? Consider the two-dimensional instance of this problem shown in Figure 9.6, which minimizes $\|(x, y)\|_p$ for $p = 2$ (pseudoinverse) and $p = 1$ (linear program). In the $p = 2$ case (a), we are minimizing $x^2 + y^2$, which has circular level sets; the optimal (x^*, y^*) subject to the constraints is in the interior of the first quadrant. In the $p = 1$ case (b), we are minimizing $|x| + |y|$, which has diamond-shaped level sets; this makes $x^* = 0$ since the outer points of the diamond align with the x and y axes, a more *sparse* solution.

More generally, the use of the norm $\|\vec{x}\|_2$ indicates that no single element x_i of $\vec{x}$ should have a large value; this regularization tends to favor vectors $\vec{x}$ with lots of small nonzero values. On the other hand, $\|\vec{x}\|_1$ does not care if a single element of $\vec{x}$ has a large value so long as the sum of all the elements' absolute values is small. As we have illustrated in the two-dimensional case, this type of regularization can produce *sparse* vectors $\vec{x}$, with elements that are exactly zero.

This type of regularization using $\|\cdot\|_1$ is fundamental in the field of *compressed sensing*, which solves underdetermined signal processing problems with the additional assumption that the output should be sparse. This assumption makes sense in many contexts where sparse solutions of $A\vec{x} = \vec{b}$ imply that many columns of A are irrelevant [34].

A minor extension of linear programming is to keep using linear inequality constraints but introduce convex quadratic terms to the objective, changing the optimization in Example 9.7 to:

$$\begin{aligned} &\text{minimize}_{\vec{x}}\ \vec{b} \cdot \vec{x} + \vec{x}^\top M \vec{x} \\ &\text{such that}\ A\vec{x} \geq \vec{c} \end{aligned}$$

Here, M is an $n \times n$ positive semidefinite matrix. With this machinery, we can provide an alternative to Tikhonov regularization from §3.1.3:

$$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2 + \alpha\|\vec{x}\|_1$$

This "lasso" regularizer also promotes sparsity in $\vec{x}$ while solving $A\vec{x} = \vec{b}$, but relaxes to the approximate case $A\vec{x} \approx \vec{b}$ in case A or $\vec{b}$ is noisy and we prefer sparsity of $\vec{x}$ over solving the system exactly [113].

9.4.2 Second-Order Cone Programming

A *second-order cone program* (SOCP) is a convex optimization problem taking the following form:

$$\begin{aligned} &\text{minimize}_{\vec{x}}\ \vec{b} \cdot \vec{x} \\ &\text{such that}\ \|A_i\vec{x} - \vec{b}_i\|_2 \leq d_i + \vec{c}_i \cdot \vec{x} \text{ for all } i = 1, \ldots, k \end{aligned}$$

Here, we have introduced matrices $A_1, \ldots, A_k$, vectors $\vec{b}_1, \ldots, \vec{b}_k$, vectors $\vec{c}_1, \ldots, \vec{c}_k$, and scalars $d_1, \ldots, d_k$ to the optimization problem. This "cone constraint" will allow us to pose additional convex optimization problems.

One non-obvious application of second-order cone programming explained in [78] appears when we wish to solve the least squares problem $A\vec{x} \approx \vec{b}$, but we do not know the elements of A exactly. For instance, A might have been constructed from data we have measured experimentally (see §3.1.2 for an example in least-squares regression).

Take $\vec{a}_i^\top$ to be the i-th row of A. Then, the least-squares problem $A\vec{x} \approx \vec{b}$ can be posed as minimizing $\sum_i (\vec{a}_i \cdot \vec{x} - b_i)^2$ over $\vec{x}$. If we do not know A exactly, however, we might allow each $\vec{a}_i$ to vary somewhat before solving least-squares. In particular, maybe we think that $\vec{a}_i$ really is an approximation of some unknown $\vec{a}_i^0$ satisfying $\|\vec{a}_i^0 - \vec{a}_i\|_2 \leq \varepsilon$ for some fixed $\varepsilon > 0$.

To make least-squares robust to this model of error, we might ask that $\vec{x}$ be chosen to thwart an adversary that chooses the worst possible $\vec{a}_i^0$. Formally, we can solve the following "minimax" problem:

$$\text{minimize}_{\vec{x}} \begin{bmatrix} \max_{\{\vec{a}_i^0\}} & \sum_i (\vec{a}_i^0 \cdot \vec{x} - b_i)^2 \\ \text{such that} & \|\vec{a}_i^0 - \vec{a}_i\| \leq \varepsilon \end{bmatrix}$$

That is, we want to choose $\vec{x}$ so that the least-squares energy with the *worst*-possible unknowns $\vec{a}_i^0$ satisfying the $\|\vec{a}_i^0 - \vec{a}_i\| \leq \varepsilon$ constraint still is small. It is far from evident that this complicated optimization problem is solvable using SOCP machinery, but after some simplification we will manage to write it in the standard SOCP form above.

If we define $\delta\vec{a}_i \equiv \vec{a}_i - \vec{a}_i^0$, then our optimization becomes:

$$\text{minimize}_{\vec{x}} \begin{bmatrix} \max_{\{\delta\vec{a}_i\}} & \sum_i (\vec{a}_i \cdot \vec{x} + \delta\vec{a}_i \cdot \vec{x} - b_i)^2 \\ \text{such that} & \|\delta\vec{a}_i\| \leq \varepsilon \end{bmatrix}$$

When maximizing over $\delta\vec{a}_i$, each term of the sum over i is independent. Hence, we can solve the maximization for one $\delta\vec{a}_i$ at a time. Peculiarly, if we maximize an absolute value rather than a sum (usually we go in the other direction!), we can find a closed-form solution to the optimization for $\delta\vec{a}_i$:

$$\begin{aligned}
\max_{\|\delta\vec{a}_i\| \leq \varepsilon} |\vec{a}_i \cdot \vec{x} + \delta a_i \cdot \vec{x} - b_i| &= \max_{\|\delta\vec{a}_i\| \leq \varepsilon} \max\{\vec{a}_i \cdot \vec{x} + \delta a_i \cdot \vec{x} - b_i, -\vec{a}_i \cdot \vec{x} - \delta a_i \cdot \vec{x} + b_i\} \\
&\qquad \text{since } |x| = \max\{x, -x\} \\
&= \max\left\{ \max_{\|\delta\vec{a}_i\| \leq \varepsilon} [\vec{a}_i \cdot \vec{x} + \delta a_i \cdot \vec{x} - b_i], \max_{\|\delta\vec{a}_i\| \leq \varepsilon} [-\vec{a}_i \cdot \vec{x} - \delta a_i \cdot \vec{x} + b_i] \right\} \\
&\qquad \text{after changing the order of the maxima} \\
&= \max\{\vec{a}_i \cdot \vec{x} + \varepsilon\|\vec{x}\|_2 - b_i, -\vec{a}_i \cdot \vec{x} + \varepsilon\|\vec{x}\|_2 + b_i\} \\
&= |\vec{a}_i \cdot \vec{x} - b_i| + \varepsilon\|\vec{x}\|_2
\end{aligned}$$

After this simplification, our optimization for $\vec{x}$ becomes:

$$\text{minimize}_{\vec{x}} \sum_i (|\vec{a}_i \cdot \vec{x} - b_i| + \varepsilon\|\vec{x}\|_2)^2$$

This minimization can be written as a second-order cone problem:

$$\begin{array}{ll} \text{minimize}_{s,\vec{t},\vec{x}} & s \\ \text{such that} & \|\vec{t}\|_2 \leq s \\ & (\vec{a}_i \cdot \vec{x} - b_i) + \varepsilon\|\vec{x}\|_2 \leq t_i \ \forall i \\ & -(\vec{a}_i \cdot \vec{x} - b_i) + \varepsilon\|\vec{x}\|_2 \leq t_i \ \forall i \end{array}$$

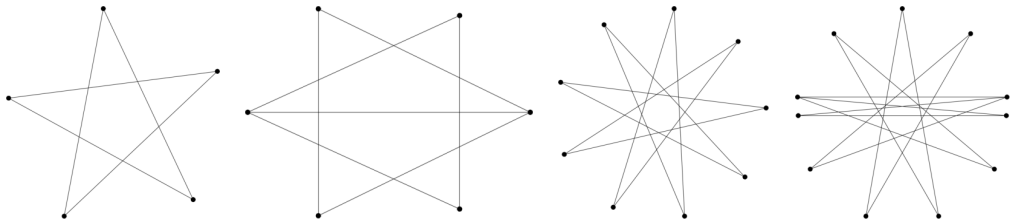


FIGURE 9.7 Examples of graphs laid out via semidefinite embedding.

In this optimization, we have introduced two extra variables s and $\vec{t}$. Since we wish to minimize s with the constraint $\|\vec{t}\|_2 \leq s$, we are effectively minimizing the norm of $\vec{t}$. The last two constraints ensure that each element of $\vec{t}$ satisfies $t_i = |\vec{a}_i \cdot \vec{x} - b_i| + \varepsilon\|\vec{x}\|_2$.

This type of regularization provides yet another variant of least-squares. In this case, rather than being robust to near-singularity of A, we have incorporated an error model directly into our formulation allowing for mistakes in the measurement of A itself. The parameter ε controls sensitivity to the elements of A in a similar fashion to the weight α of Tikhonov or L_1 regularization.

9.4.3 Semidefinite Programming

Suppose A and B are $n \times n$ positive semidefinite matrices; we will notate this as $A, B \succeq 0$. Take $t \in [0, 1]$. Then, for any $\vec{x} \in \mathbb{R}^n$ we have:

$$\vec{x}^\top(tA + (1-t)B)\vec{x} = t\vec{x}^\top A\vec{x} + (1-t)\vec{x}^\top B\vec{x} \geq 0,$$

where the inequality holds by semidefiniteness of A and B. This proof verifies a surprisingly useful fact:

The set of positive semidefinite matrices is convex.

Hence, if we are solving optimization problems for a matrix A, we safely can add constraints $A \succeq 0$ without affecting convexity.

Algorithms for *semidefinite programming* optimize convex objectives with the ability to add constraints that matrix-valued variables must be positive (or negative) semidefinite. More generally, semidefinite programming machinery can include *linear matrix inequality* (LMI) constraints of the form:

$$x_1A_1 + x_2A_2 + \cdots + x_kA_k \succeq 0,$$

where $\vec{x} \in \mathbb{R}^k$ is an optimization variable and the matrices A_i are fixed.

As an example of semidefinite programming, we will sketch a technique known as *semidefinite embedding* from graph layout and manifold learning [124]. Suppose we are given a graph (V, E) consisting of a set of vertices $V = \{v_1, \ldots, v_k\}$ and a set of edges $E \subseteq V \times V$. For some fixed n, the semidefinite embedding method computes positions $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$ for the vertices, so that vertices connected by edges are nearby in the embedding with respect to Euclidean distance $\|\cdot\|_2$; some examples are shown in Figure 9.7.

If we already have computed $\vec{x}_1, \ldots, \vec{x}_k$, we can construct a "Gram matrix" $G \in \mathbb{R}^{k\times k}$ satisfying $G_{ij} = \vec{x}_i \cdot \vec{x}_j$. G is a matrix of inner products and hence is symmetric and positive

semidefinite. We can measure the squared distance from $\vec{x}_i$ to $\vec{x}_j$ using G:

$$\begin{aligned}\|\vec{x}_i - \vec{x}_j\|_2^2 &= (\vec{x}_i - \vec{x}_j) \cdot (\vec{x}_i - \vec{x}_j) \\ &= \|\vec{x}_i\|_2^2 - 2\vec{x}_i \cdot \vec{x}_j + \|\vec{x}_j\|_2^2 \\ &= G_{ii} - 2G_{ij} + G_{jj}\end{aligned}$$

Similarly, suppose we wish the center of mass $\frac{1}{k}\sum_i \vec{x}_i$ to be $\vec{0}$, since shifting the embedding of the graph does not have a significant effect on its layout. We alternatively can write $\|\sum_i \vec{x}_i\|_2^2 = 0$, allowing us to express this condition in terms of G:

$$0 = \left\|\sum_i \vec{x}_i\right\|_2^2 = \left(\sum_i \vec{x}_i\right) \cdot \left(\sum_i \vec{x}_i\right) = \sum_{ij} \vec{x}_i \cdot \vec{x}_j = \sum_{ij} G_{ij}$$

Finally, we might wish that our embedding of the graph is relatively compact or small. One way to do this would be to minimize $\sum_i \|\vec{x}_i\|^2 = \sum_i G_{ii} = \text{Tr}(G)$.

The semidefinite embedding technique turns these observations on their head, optimizing for the Gram matrix G rather than the positions $\vec{x}_i$ of the vertices. In particular, consider the following optimization problem:

$$\begin{aligned}\text{minimize}_{G \in \mathbb{R}^{k \times k}} \quad & \text{Tr}(G) \\ \text{such that} \quad & G = G^\top \\ & G \succeq 0 \\ & G_{ii} - 2G_{ij} + G_{jj} = 1 \; \forall (v_i, v_j) \in E \\ & \textstyle\sum_{ij} G_{ij} = 0\end{aligned}$$

This optimization for G is motivated as follows:

- The objective asks that the embedding of the graph is compact by minimizing the sum of squared norms of each $\vec{x}_i$.
- The first two constraints require that the Gram matrix is symmetric and positive definite.
- The third constraint asks that the embedding of any two adjacent vertices in the graph have distance one.
- The final constraint centers the embedding about the origin.

We can use semidefinite programming to solve this optimization problem for G. Then, since G is symmetric and positive semidefinite, we can use the Cholesky factorization (§3.2.1) or the eigenvector decomposition (§5.2) of G to write $G = X^\top X$ for some matrix $X \in \mathbb{R}^{k \times k}$. Based on our discussion above, the columns of X are an embedding of the vertices of the graph into $\mathbb{R}^k$ where all the edges in the graph have length one, the center of mass is the origin, and the total square norm of the positions is minimized *exactly*.

We set out to embed the graph into $\mathbb{R}^n$ rather than $\mathbb{R}^k$, and generally $n \leq k$. To compute a lower-dimensional embedding that *approximately* satisfies the constraints above, we can decompose $G = X^\top X$ using its eigenvectors; then, we remove $k - n$ eigenvectors with eigenvalues closest to zero; this operation is exactly the low-rank approximation of G via SVD given in §6.2.2. This final step provides the desired embedding into $\mathbb{R}^n$.

A legitimate question about the semidefinite embedding is how the optimization for G interacts with its low-rank eigenvector approximation applied in post-processing. In fact, in

many well-known cases the solution of semidefinite optimizations like the one above yield *low-rank* or nearly low-rank matrices whose lower-dimensional approximations are close to the original input; a formalized version of this observation justifies the final step. We already explored such a justification in exercise 6.7, since the nuclear norm of a symmetric positive semidefinite matrix is its trace.

9.4.4 Integer Programs and Relaxations

Our final application of convex optimization is—surprisingly—to a class of *highly* non-convex problems: Ones with integer variables. In particular, an *integer program* is an optimization in which one or more variables is constrained to be an integer rather than a real number. Within this class, two well-known subproblems are *mixed-integer programming*, in which some variables are continuous while others are integers, and *zero-one programming*, where the variables take boolean values in $\{0, 1\}$.

Example 9.10 (3-SAT)**.** We can define the following operations from boolean algebra for binary variables $U, V \in \{0, 1\}$:

U	V	$\neg U$ ("not U")	$\neg V$ ("not V")	$U \wedge V$ ("U and V")	$U \vee V$ ("U or V")
0	0	1	1	0	0
0	1	1	0	0	1
1	0	0	1	0	1
1	1	0	0	1	1

We can convert boolean satisfiability problems into integer programs using a few steps. For example, we can express the "not" operation algebraically using $\neg U = 1 - U$. Similarly, suppose we wish to find U, V satisfying $(U \vee \neg V) \wedge (\neg U \vee V)$. Then, U and V as integers satisfy the following constraints:

$$\begin{aligned} U + (1 - V) &\geq 1 && (U \vee \neg V) \\ (1 - U) + V &\geq 1 && (\neg U \vee V) \\ U, V &\in \mathbb{Z} && \text{(integer constraint)} \\ 0 \leq U, V &\leq 1 && \text{(boolean variables)} \end{aligned}$$

As demonstrated in Example 9.10, integer programs encode a wide class of discrete problems, including many that are known to be NP-hard. For this reason, we cannot expect to solve them exactly with convex optimization machinery, since doing so would settle the long-standing question of theoretical computer science by showing "$P = NP$." We can, however, use convex optimization to find approximate solutions to integer programming problems.

In particular, if we write a discrete problem like Example 9.10 as an optimization, we can *relax* the constraint keeping variables in $\mathbb{Z}$ and allow them to be in $\mathbb{R}$ instead. Such a relaxation can yield invalid solutions, e.g. boolean variables that take on values like 0.75. So, after solving the relaxed problem, one of many strategies can be used to generate an integer approximation of the solution. For example, non-integral variables can be *rounded* to the closest integer, at the risk of generating outputs that are both suboptimal and violate the constraints. Alternatively, a slower but potentially more effective method iteratively rounds one variable at a time, adds a constraint fixing the value of that variable, and re-optimizes the objective subject to the new constraint.

Many difficult discrete problems can be reduced to integer programming, from satis-

fiability problems like the one in Example 9.10 to the traveling salesman problem. These reductions should indicate that the design of effective integer programming techniques is challenging even in the approximate case. State-of-the-art methods for integer programming, however, are fairly effective for a large class of problems, providing a remarkably general piece of machinery for producing approximate solutions to problems for which it may be difficult or impossible to design a discrete solution algorithm. Many open research problems involve designing effective integer programming methods and understanding potential relaxations; this work provides a valuable and attractive link between continuous and discrete mathematics and modeling.

9.5 EXERCISES

9.1 Prove the following statement from §9.4: If f is a convex function, the set $\{\vec{x} : f(\vec{x}) \leq c\}$ is convex.

9.2 The standard deviation of k values $x_1, \ldots, x_k$ is

$$\sigma(x_1, \ldots, x_k) \equiv \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \mu)^2},$$

where $\mu \equiv \frac{1}{k}\sum_i x_i$. Show that σ is a convex function of $x_1, \ldots, x_k$.

9.3 Some properties of second-order cone programming:

(a) Show that the *Lorentz cone* $\{\vec{x} \in \mathbb{R}^n, c \in \mathbb{R} : \|\vec{x}\|_2 \leq c\}$ is convex.

(b) Use this fact to show that the second-order cone program in §9.4.2 is convex.

(c) Show that second-order cone programming can be used to solve linear programs.

9.4 In this problem we will study *linear programming* in more detail.

(a) A linear program in "standard form" is given by:

$$\begin{array}{ll} \text{minimize}_{\vec{x}} & \vec{c}^\top \vec{x} \\ \text{such that} & A\vec{x} = \vec{b} \\ & \vec{x} \geq \vec{0} \end{array}$$

Here, the optimization is over $\vec{x} \in \mathbb{R}^n$; the remaining variables are constants $A \in \mathbb{R}^{m \times n}$, $\vec{b} \in \mathbb{R}^m$, and $\vec{c} \in \mathbb{R}^n$. Find the KKT conditions of this system.

(b) Suppose we add a constraint of the form $\vec{v}^\top \vec{x} \leq d$ for some fixed $\vec{v} \in \mathbb{R}^n$ and $d \in \mathbb{R}$. Explain how such a constraint can be added while keeping a linear program in standard form.

(c) The "dual" of this linear program is another optimization:

$$\begin{array}{ll} \text{maximize}_{\vec{y}} & \vec{b}^\top \vec{y} \\ \text{such that} & A^\top \vec{y} \leq \vec{c} \end{array}$$

Assuming that the primal and dual have exactly one stationary point, show that the optimal value of the primal and dual objectives coincide.

Hint: Show that the KKT multipliers of one problem can be used to solve the other.
Note: This property is called "strict duality." The famous simplex algorithm for solving linear programs maintains estimates of $\vec{x}$ and $\vec{y}$, terminating when $\vec{c}^\top \vec{x}^* - \vec{b}^\top \vec{y}^* = 0$.

9.5 Suppose we take a grayscale photograph of size $n \times m$ and represent it as a vector $\vec{v} \in \mathbb{R}^{nm}$ of values in $[0, 1]$. We used the wrong lens, however, and our photo is blurry! We wish to use *deconvolution* machinery to undo this effect.

(a) Find the KKT conditions for the following optimization problem:

$$\begin{array}{rl} \text{minimize}_{\vec{x} \in \mathbb{R}^{nm}} & \|A\vec{x} - \vec{b}\|_2^2 \\ \text{such that} & 0 \leq x_i \leq 1 \ \forall i \in \{1, \ldots, nm\} \end{array}$$

(b) Suppose we are given a matrix $G \in \mathbb{R}^{nm \times nm}$ taking sharp images to blurry ones. Propose an optimization in the form of (a) for recovering a sharp image from our blurry $\vec{v}$.

(c) We do not know the operator G, making the model in (b) difficult to use. Suppose, however, that for each $r \geq 0$ we can write a matrix $G_r \in \mathbb{R}^{nm \times nm}$ approximating a blur with radius r. Using the same camera, we now take k *pairs* of photos $(\vec{v}_1, \vec{w}_1), \ldots, (\vec{v}_k, \vec{w}_k)$, where $\vec{v}_i$ and $\vec{w}_i$ are of the same scene but $\vec{v}_i$ is blurry (taken using the same lens as our original bad photo) and $\vec{w}_i$ is sharp. Propose a nonlinear optimization for approximating r using this data.

9.6[DH] ("Fenchel duality," adapted from [10]) Let $f(\vec{x})$ be a convex function on $\mathbb{R}^n$ that is *proper*. This means that f accepts vectors from $\mathbb{R}^n$ or whose coordinates may (individually) be $\pm\infty$ and returns a real scalar in $\mathbb{R} \cup \{\infty\}$ with at least one $f(\vec{x}_0)$ taking a non-infinite value. Under these assumptions, the *Fenchel dual* of f at $\vec{y} \in \mathbb{R}^n$ is defined to be the function

$$f^*(\vec{y}) \equiv \sup_{\vec{x} \in \mathbb{R}^n} (\vec{x} \cdot \vec{y} - f(\vec{x})).$$

Fenchel duals are used to study properties of convex optimization problems in theory and practice.

(a) Show that f^* is convex.

(b) Derive the *Fenchel-Young inequality*:

$$f(\vec{x}) + f^*(\vec{y}) \geq \vec{x} \cdot \vec{y}.$$

(c) The *indicator function* of a subset $A \in \mathbb{R}^n$ is given by

$$\chi_A(\vec{x}) \equiv \begin{cases} 0 & \text{if } \vec{x} \in A \\ \infty & \text{otherwise} \end{cases}$$

With this definition in mind, determine the Fenchel dual of $f(\vec{x}) = \vec{c} \cdot \vec{x}$, where $\vec{c} \in \mathbb{R}^n$.

(d) What is the Fenchel dual of the linear function $f(x) = ax + b$?

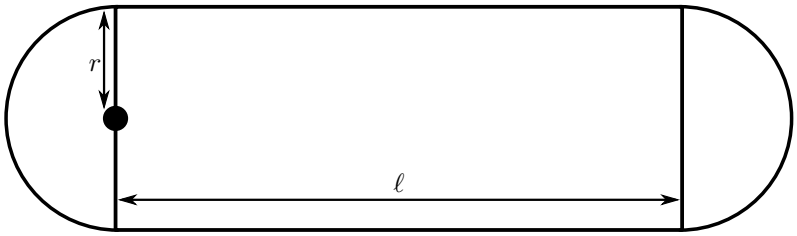


FIGURE 9.8 Notation for problem 9.7.

(e) Show that $f(\vec{x}) = \frac{1}{2}\|\vec{x}\|_2^2$ is *self-dual*, meaning $f = f^*$.

(f) Suppose $p, q \in (1, \infty)$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Show that the Fenchel dual of $f(x) = \frac{1}{p}|x|^p$ is $f^*(y) = \frac{1}{q}|y|^q$. Use this result along with previous parts of this problem to derive Hölder's inequality:

$$\sum_k |u_k v_k| \leq \left(\sum_k |u_k|^p\right)^{1/p} \left(\sum_k |v_k|^q\right)^{1/q}$$

for all $\vec{u}, \vec{v} \in \mathbb{R}^n$.

9.7[SC] A monomial is a function of the form $f(\vec{x}) = cx_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$, where each $a_i \in \mathbb{N} \cup \{0\}$. We additionally define a *posynomial* as a sum of monomials with positive coefficients:

$$f(\vec{x}) = \sum_{k=1}^{K} c_k x_1^{a_{k_1}} x_2^{a_{k_2}} \cdots x_n^{a_{k_n}},$$

where $c_k \geq 0$ for all k.

Geometric programs are optimizations taking the following form:

$$\begin{array}{rl} \text{minimize}_{\vec{x}} & f_0(\vec{x}) \\ \text{such that} & f_i(\vec{x}) \leq 1 \, \forall i \in \{1, \ldots, m\} \\ & g_i(\vec{x}) = 1 \, \forall i \in \{1, \ldots, p\}, \end{array}$$

where the functions f_i are posynomials and the functions g_i are monomials.

(a) Suppose you are designing a slow-dissolving medicinal capsule. The capsule looks like a cylinder with hemispherical ends, illustrated in Figure 9.8. To ensure that the capsule dissolves slowly, you need to minimize its surface area.
The cylindrical portion of the capsule must have volume larger than or equal to V to ensure that it can hold the proper amount of medicine. Also, because the capsule is manufactured as two halves that slide together, to ensure that the capsule will not break, the length ℓ of its cylindrical portion must be at least $\ell_{\min}$. Finally, due to packaging limitations the total length of the capsule must be no larger than C.
Write the corresponding minimization problem and argue that it is a geometric program.

(b) Transform the problem from part 9.7a into a convex programming problem.
Hint: Consider the substitution $y_i = \log x_i$.

9.8 The *cardinality function* $\|\cdot\|_0$ computes the number of nonzero elements of $\vec{x} \in \mathbb{R}^n$:

$$\|\vec{x}\|_0 = \sum_{i=1}^{n} \begin{cases} 1 & x_i \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that $\|\cdot\|_0$ is *not* a norm on $\mathbb{R}^n$, but that it is connected to L_p norms by the relationship

$$\|\vec{x}\|_0 = \lim_{p \to 0} \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

(b) Suppose we wish to solve an underdetermined system of equations $A\vec{x} = \vec{b}$. One alternative to SVD-based approaches or Tikhonov regularizations is *cardinality minimization*:

$$\begin{array}{rl} \min_{\vec{x} \in \mathbb{R}^n} & \|\vec{x}\|_0 \\ \text{such that} & A\vec{x} = \vec{b} \\ & \|\vec{x}\|_\infty \leq R. \end{array}$$

Rewrite this optimization in the form

$$\begin{array}{rl} \min_{\vec{x}, \vec{z}} & \|\vec{z}\|_1 \\ \text{such that} & \vec{z} \in \{0,1\}^n \\ & \vec{x}, \vec{z} \in \mathcal{C}, \end{array}$$

where $\mathcal{C}$ is some convex set [13].

(c) Show that relaxing the constraint $\vec{z} \in \{0,1\}^n\}$ to $\vec{z} \in [0,1]^n$ lower-bounds the original problem. Propose a heuristic for solving the $\{0,1\}$ problem based on this relaxation.

9.9 ("Grasping force optimization;" adapted from [78]) Suppose we are writing code to control a robot hand with n fingers grasping a rigid object. Each finger i is controlled by a motor that outputs torque t_i.

The force $\vec{F}_i$ imparted by each finger onto the object can be decomposed into two orthogonal parts $\vec{F}_i = \vec{F}_{ni} + \vec{F}_{si}$, a normal force $\vec{F}_{ni}$ and a tangential friction force $\vec{F}_{si}$:

$$\textbf{Normal force: } \vec{F}_{ni} = c_i t_i \vec{v}_i = (\vec{v}_i^\top \vec{F}_i)\vec{v}_i$$
$$\textbf{Friction force: } \vec{F}_{si} = (I_{3\times 3} - \vec{v}_i \vec{v}_i^\top)\vec{F}_i, \text{ where } \|\vec{F}_{si}\|_2 \leq \mu \|F_{ni}\|_2$$

Here, $\vec{v}_i$ is a (fixed) unit vector normal to the surface at the point of contact of finger i. The value c_i is a constant associated with finger i. Additionally, the object experiences a gravitational force in the downward direction given by $\vec{F}_g = m\vec{g}$.

For the object to be grasped firmly in place, the sum of the forces exerted by all fingers must be $\vec{0}$. Show how to minimize the total torque outputted by the motors while firmly grasping the object using a second-order cone program.

9.10 Show that when $\vec{c}_i = \vec{0}$ for all i in the second-order cone program of §9.4.2, the optimization problem can be solved as a convex quadratic program with quadratic constraints.

9.11 (Suggested by Q. Huang) Suppose we know

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & x \\ 1 & x & 1 \end{pmatrix} \succeq 0.$$

What can we say about x?

9.12[DH] We say that $A \in \mathbb{R}^{p \times p}$ is *unimodular* if its determinant is ± 1. More generally, $M \in \mathbb{R}^{m \times n}$ is *totally unimodular* if and only if all of its invertible submatrices are unimodular. Suppose we are given a linear program whose constraints can be written in the form $M\vec{x} \leq \vec{b}$, where $\vec{b}$ is a vector of integers and M is totally unimodular. Show that in this case the linear program admits an integral solution.

9.13[SC] We can modify the gradient descent algorithm for minimizing $f(\vec{x})$ to account for linear equality constraints $A\vec{x} = \vec{b}$.

(a) Assuming we choose $\vec{x}_0$ satisfying the equality constraint, propose a modification to gradient descent so that each iterate $\vec{x}_k$ satisfies $A\vec{x}_k = \vec{b}$.
Hint: The gradient $\nabla f(\vec{x})$ may point in a direction that could violate the constraint.

(b) Briefly justify why the modified gradient descent algorithm should reach a local minimum of the constrained optimization problem.

(c) Suppose rather than $A\vec{x} = \vec{b}$ we have a nonlinear constraint $g(\vec{x}) = \vec{0}$. Propose a modification of your strategy from 9.13a maintaining this new constraint approximately. How is the modification affected by the choice of step sizes?

9.14 Show that linear programming and second-order cone programming are special cases of semidefinite programming.

CHAPTER 10

Iterative Linear Solvers

CONTENTS

10.1 Gradient Descent 200
10.1.1 Gradient Descent for Linear Systems 200
10.1.2 Convergence 201
10.2 Conjugate Gradients 203
10.2.1 Motivation 204
10.2.2 Suboptimality of Gradient Descent 205
10.2.3 Generating A-Conjugate Directions 207
10.2.4 Formulating the Conjugate Gradients Algorithm 208
10.2.5 Convergence and Stopping Conditions 211
10.3 Preconditioning 211
10.3.1 CG with Preconditioning 212
10.3.2 Common Preconditioners 213
10.4 Other Iterative Schemes 214

In the previous two chapters, we developed strategies for minimizing a function $f(\vec{x})$ with or without constraints on $\vec{x}$. In doing so, we relaxed our viewpoint from numerical linear algebra that we must find an *exact* solution to a system of equations and instead designed iterative schemes that successively produce better approximations of the minimum of a function. Even if we never find the position $\vec{x}^*$ of a local minimum exactly, such methods generate $\vec{x}_k$ with smaller and smaller $f(\vec{x}_k)$, in many cases getting arbitrarily close to the desired optimal output.

We now revisit our favorite problem from numerical linear algebra, solving $A\vec{x} = \vec{b}$ for $\vec{x}$, but apply an *iterative* approach rather than expecting to find a solution in closed form. This adjustment reveals a new class of linear system solvers that can find reliable approximations of $\vec{x}$ in remarkably few iterations. To formulate these methods, we will view solving $A\vec{x} = \vec{b}$ not as a system of equations but rather as a minimization problem, e.g. on energies like $\|A\vec{x} - \vec{b}\|_2^2$.

Why bother deriving yet another class of linear system solvers? So far, most of our direct solvers require us to represent A as a full $n \times n$ matrix, and algorithms such as LU, QR, or Cholesky factorization all take around $O(n^3)$ time. Two cases motivate the need for iterative linear solvers:

- When A is sparse, Gaussian elimination tends to induce *fill*, meaning that even if A contains $O(n)$ nonzero values, intermediate steps of elimination may fill in the remaining $O(n^2)$ empty positions. Storing a matrix in sparse format dramatically reduces the space it takes in memory, but fill during elimination rapidly can cancel out these savings. Contrastingly, the algorithms in this chapter require only that you can *apply* A to vectors (that is, that we can find the product $A\vec{v}$ for any $\vec{v}$), which does

not induce fill and can be carried out in time proportional to the number of nonzeros in a sparse matrix.

- We may wish to defeat the $O(n^3)$ runtime of standard matrix factorization techniques. In particular, if an iterative scheme can uncover a fairly, if not completely, accurate solution to $A\vec{x} = \vec{b}$ in a few steps, we may halt the method early in favor of speed over accuracy of the output.

Also, Newton's method and other nonlinear optimization methods we have discussed require solving a linear system in each iteration. Formulating the fastest possible solver can make a considerable difference when implementing large-scale optimization methods that require one or more linear solves per iteration. In this case an inaccurate but fast linear solve may be acceptable, since it feeds into a larger iterative technique anyway.

Although our discussion in this chapter benefits from intuition and formalism developed in previous chapters, our approach to deriving iterative linear methods owes much to the classic and more extended treatment in [104].

10.1 GRADIENT DESCENT

We will focus our discussion on solving $A\vec{x} = \vec{b}$ where A has three properties:

1. $A \in \mathbb{R}^{n \times n}$ is square
2. A is symmetric, that is, $A^\top = A$
3. A is positive definite, that is, for all $\vec{x} \neq \vec{0}$, $\vec{x}^\top A\vec{x} > 0$

Toward the end of this chapter we will relax these assumptions. For now, we can replace $A\vec{x} = \vec{b}$—at least when A is invertible or overdetermined—with the normal equations $A^\top A\vec{x} = A^\top\vec{b}$ to satisfy these criteria, although as in §4.1 this substitution can create conditioning issues.

10.1.1 Gradient Descent for Linear Systems

In this case, solutions of $A\vec{x} = \vec{b}$ are minima of the function $f(\vec{x})$ given by the *quadratic form*

$$f(\vec{x}) \equiv \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c$$

for any $c \in \mathbb{R}$. In particular, when A is symmetric, taking the derivative of f shows

$$\nabla f(\vec{x}) = A\vec{x} - \vec{b},$$

and setting $\nabla f(\vec{x}) = \vec{0}$ yields the desired result.

Solving $\nabla f(\vec{x}) = \vec{0}$ directly amounts to performing Gaussian elimination on A. Instead, however, suppose we apply gradient descent to this minimization. Recall the basic gradient descent algorithm:

1. Compute the search direction $\vec{d}_k \equiv -\nabla f(\vec{x}_{k-1}) = \vec{b} - A\vec{x}_{k-1}$.
2. Define $\vec{x}_k \equiv \vec{x}_{k-1} + \alpha_k\vec{d}_k$, where α_k is chosen such that $f(\vec{x}_k) < f(\vec{x}_{k-1})$

```
function LINEAR-GRADIENT-DESCENT(A, b⃗)
  x⃗ ← 0⃗
  for k ← 1, 2, 3, ...
    d⃗ ← b⃗ − Ax⃗                          ▷ Search direction is residual
    α ← ‖d⃗‖₂² / (d⃗ᵀ A d⃗)                ▷ Line search formula
    x⃗ ← x⃗ + αd⃗                          ▷ Update solution vector x⃗
```

FIGURE 10.1 Gradient descent algorithm for solving $A\vec{x} = \vec{b}$ for symmetric and positive definite A, by iteratively decreasing the energy $f(\vec{x}) = \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c$.

For a generic function f, deciding on the value of α_k can be a difficult one-dimensional "line search" problem, boiling down to minimizing $f(\vec{x}_{k-1} + \alpha_k\vec{d}_k)$ as a function of a single variable $\alpha_k \geq 0$. For the quadratic form $f(\vec{x}) = \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c$, however, we can choose α_k optimally using a closed-form formula. To do so, define

$$\begin{aligned}
g(\alpha) &\equiv f(\vec{x} + \alpha\vec{d}) \\
&= \frac{1}{2}(\vec{x} + \alpha\vec{d})^\top A(\vec{x} + \alpha\vec{d}) - \vec{b}^\top(\vec{x} + \alpha\vec{d}) + c \\
&= \frac{1}{2}(\vec{x}^\top A\vec{x} + 2\alpha\vec{x}^\top A\vec{d} + \alpha^2\vec{d}^\top A\vec{d}) - \vec{b}^\top\vec{x} - \alpha\vec{b}^\top\vec{d} + c \text{ after expanding the product} \\
&= \frac{1}{2}\alpha^2\vec{d}^\top A\vec{d} + \alpha(\vec{x}^\top A\vec{d} - \vec{b}^\top\vec{d}) + \text{const.} \\
\implies \frac{dg}{d\alpha}(\alpha) &= \alpha\vec{d}^\top A\vec{d} + \vec{d}^\top(A\vec{x} - \vec{b}) \text{ by symmetry of } A
\end{aligned}$$

With this simplification in place, if we wish to minimize g with respect to α, we solve $dg/d\alpha = 0$ to find

$$\alpha = \frac{\vec{d}^\top(\vec{b} - A\vec{x})}{\vec{d}^\top A\vec{d}}$$

For gradient descent we chose $\vec{d}_k = \vec{b} - A\vec{x}_k$, so α_k takes the form:

$$\alpha_k = \frac{\|\vec{d}_k\|_2^2}{\vec{d}_k^\top A\vec{d}_k}$$

In the end, our formula for line search yields the iterative gradient descent scheme for solving $A\vec{x} = \vec{b}$ shown in Figure 10.1.

10.1.2 Convergence

By construction, gradient descent decreases $f(\vec{x}_k)$ as $k \to \infty$. Even so, we have not shown that the algorithm approaches the minimum $f(\vec{x}_k)$, and we have not been able to characterize how many iterations we should run to reach a reasonable level of confidence that $A\vec{x}_k \approx \vec{b}$. One way to understand the convergence of the gradient descent algorithm for our choice of f is to examine the change in backward error from iteration to iteration; we will follow the argument in [35] and elsewhere.

Suppose $\vec{x}^*$ satisfies $A\vec{x}^* = \vec{b}$ exactly. Then, the backward error in iteration k is given by:

$$R_k \equiv \frac{f(\vec{x}_k) - f(\vec{x}^*)}{f(\vec{x}_{k-1}) - f(\vec{x}^*)}$$

Bounding $R_k < \beta < 1$ for some fixed β (possibly depending on A) would imply $R_k \to 0$ as $k \to \infty$; in this case, $f(\vec{x}_k) \to f(\vec{x}^*)$, showing that the gradient descent algorithm converges.

For convenience, we can expand $f(\vec{x}_k)$:

$$\begin{aligned}
f(\vec{x}_k) &= f(\vec{x}_{k-1} + \alpha_k \vec{d}_k) \text{ by our iterative scheme} \\
&= \frac{1}{2}(\vec{x}_{k-1} + \alpha_k \vec{d}_k)^\top A(\vec{x}_{k-1} + \alpha_k \vec{d}_k) - \vec{b}^\top(\vec{x}_{k-1} + \alpha_k \vec{d}_k) + c \\
&= f(\vec{x}_{k-1}) + \alpha_k \vec{d}_k^\top A\vec{x}_{k-1} + \frac{1}{2}\alpha_k^2 \vec{d}_k^\top A\vec{d}_k - \alpha_k \vec{b}^\top \vec{d}_k \text{ by definition of } f \\
&= f(\vec{x}_{k-1}) + \alpha_k \vec{d}_k^\top(\vec{b} - \vec{d}_k) + \frac{1}{2}\alpha_k^2 \vec{d}_k^\top A\vec{d}_k - \alpha_k \vec{b}^\top \vec{d}_k \text{ since } \vec{d}_k = \vec{b} - A\vec{x}_{k-1} \\
&= f(\vec{x}_{k-1}) - \alpha_k \vec{d}_k^\top \vec{d}_k + \frac{1}{2}\alpha_k^2 \vec{d}_k^\top A\vec{d}_k \text{ since the remaining terms cancel} \\
&= f(\vec{x}_{k-1}) - \frac{\vec{d}_k^\top \vec{d}_k}{\vec{d}_k^\top A\vec{d}_k}(\vec{d}_k^\top \vec{d}_k) + \frac{1}{2}\left(\frac{\vec{d}_k^\top \vec{d}_k}{\vec{d}_k^\top A\vec{d}_k}\right)^2 \vec{d}_k^\top A\vec{d}_k \text{ by definition of } \alpha_k \\
&= f(\vec{x}_{k-1}) - \frac{(\vec{d}_k^\top \vec{d}_k)^2}{2\vec{d}_k^\top A\vec{d}_k}
\end{aligned}$$

We can use this formula to find an alternative expression for the backward error R_k:

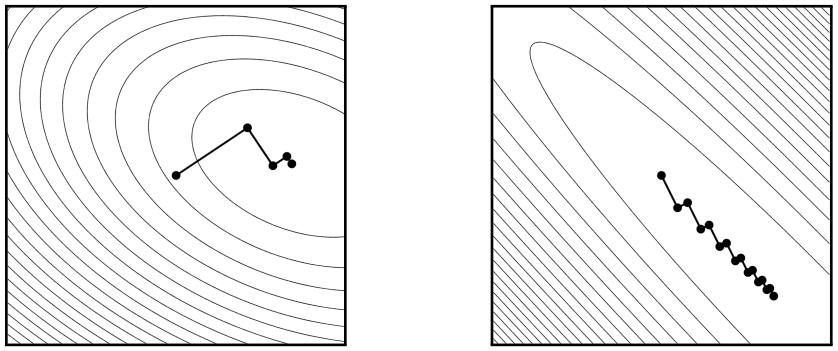
$$\begin{aligned}
R_k &= \frac{f(\vec{x}_{k-1}) - \frac{(\vec{d}_k^\top \vec{d}_k)^2}{2\vec{d}_k^\top A\vec{d}_k} - f(\vec{x}^*)}{f(\vec{x}_{k-1}) - f(\vec{x}^*)} \text{ by our formula for } f(\vec{x}_k) \\
&= 1 - \frac{(\vec{d}_k^\top \vec{d}_k)^2}{2\vec{d}_k^\top A\vec{d}_k(f(\vec{x}_{k-1}) - f(\vec{x}^*))}
\end{aligned}$$

To simplify the difference in the denominator, we can use $\vec{x}^* = A^{-1}\vec{b}$ to write:

$$\begin{aligned}
f(\vec{x}_{k-1}) - f(\vec{x}^*) &= \left[\frac{1}{2}\vec{x}_{k-1}^\top A\vec{x}_{k-1} - \vec{b}^\top \vec{x}_{k-1} + c\right] - \left[\frac{1}{2}(\vec{x}^*)^\top \vec{b} - \vec{b}^\top \vec{x}^* + c\right] \\
&= \frac{1}{2}\vec{x}_{k-1}^\top A\vec{x}_{k-1} - \vec{b}^\top \vec{x}_{k-1} - \frac{1}{2}\vec{b}^\top A^{-1}\vec{b} \text{ by our expression for } \vec{x}^* \\
&= \frac{1}{2}(A\vec{x}_{k-1} - \vec{b})^\top A^{-1}(A\vec{x}_{k-1} - \vec{b}) \text{ by symmetry of } A \\
&= \frac{1}{2}\vec{d}_k^\top A^{-1}\vec{d}_k \text{ by definition of } \vec{d}_k
\end{aligned}$$

Plugging this expression into our simplified formula for R_k shows:

$$\begin{aligned}
R_k &= 1 - \frac{(\vec{d}_k^\top \vec{d}_k)^2}{\vec{d}_k^\top A\vec{d}_k \cdot \vec{d}_k^\top A^{-1}\vec{d}_k} \\
&= 1 - \frac{\vec{d}_k^\top \vec{d}_k}{\vec{d}_k^\top A\vec{d}_k} \cdot \frac{\vec{d}_k^\top \vec{d}_k}{\vec{d}_k^\top A^{-1}\vec{d}_k} \\
&\le 1 - \left(\min_{\|\vec{d}\|=1} \frac{1}{\vec{d}^\top A\vec{d}}\right)\left(\min_{\|\vec{d}\|=1} \frac{1}{\vec{d}^\top A^{-1}\vec{d}}\right) \text{ since this makes the second term smaller} \\
&= 1 - \left(\max_{\|\vec{d}\|=1} \vec{d}^\top A\vec{d}\right)^{-1}\left(\max_{\|\vec{d}\|=1} \vec{d}^\top A^{-1}\vec{d}\right)^{-1}
\end{aligned}$$



Well-conditioned A Poorly-conditioned A

FIGURE 10.2 Gradient descent starting from the origin $\vec{0}$ (at the center) on $f(\vec{x}) = \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c$ for two choices of A. Each figure shows level sets of $f(\vec{x})$ as well as iterates of gradient descent connected by line segments.

$$= 1 - \frac{\sigma_{\min}}{\sigma_{\max}} \text{ where } \sigma_{\min}, \sigma_{\max} \text{ are the minimum/maximum singular values of } A$$
$$= 1 - \frac{1}{\operatorname{cond} A}$$

It took a considerable amount of algebra, but we proved an important fact:

Convergence of gradient descent on f depends on the conditioning of A.

That is, the better conditioned A is, the faster gradient descent will converge. Additionally, since $\operatorname{cond} A \geq 1$, we know that gradient descent converges *unconditionally* to $\vec{x}^*$, although convergence can be slow when A is poorly-conditioned.

Figure 10.2 illustrates behavior of gradient descent for well- and poorly-conditioned matrices A. When the eigenvalues of A have a wide spread, A is poorly-conditioned and gradient descent struggles to find the minimum of our quadratic function f, zig-zagging along the energy landscape.

10.2 CONJUGATE GRADIENTS

Solving $A\vec{x} = \vec{b}$ for dense $A \in \mathbb{R}^{n\times n}$ takes $O(n^3)$ time using Gaussian elimination. Reexamining gradient descent from §10.1.1 above, we see that in the dense case each iteration takes $O(n^2)$ time, since we must compute matrix-vector products between A and $\vec{x}_{k-1}, \vec{d}_k$. So, if gradient descent takes more than n iterations, from a timing standpoint we might as well have used Gaussian elimination, which would have recovered the *exact* solution in the same amount of time. Unfortunately, gradient descent may never reach the exact solution $\vec{x}^*$ in a finite number of iterations, and in poorly-conditioned cases it can take a huge number of iterations to approximate $\vec{x}^*$ well.

For this reason, we will design the *conjugate gradients* (CG) algorithm, which is *guaranteed* to converge in at most n steps, preserving $O(n^3)$ worst-case timing for solving linear systems. We also will find that this algorithm exhibits better convergence properties overall, often making it preferable to gradient descent even if we do not run it to completion.

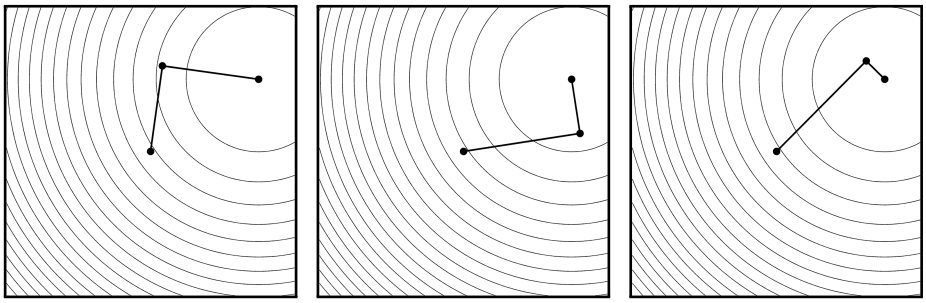


FIGURE 10.3 Searching along *any* two orthogonal directions minimizes $\bar{f}(\vec{y}) = \|\vec{y} - \vec{y}^*\|_2^2$ over $\vec{y} \in \mathbb{R}^2$. Each example in this figure has the same starting point but searches along a different pair of orthogonal directions; in the end they all reach the same optimal point.

10.2.1 Motivation

Our derivation of the conjugate gradients algorithm is motivated by writing the energy functional $f(\vec{x})$ in an alternative form. Suppose we knew the solution $\vec{x}^*$ to $A\vec{x}^* = \vec{b}$. Then, we can write:

$$\begin{aligned}
f(\vec{x}) &= \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c \text{ by definition} \\
&= \frac{1}{2}(\vec{x} - \vec{x}^*)^\top A(\vec{x} - \vec{x}^*) + \vec{x}^\top A\vec{x}^* - \frac{1}{2}(\vec{x}^*)^\top A\vec{x}^* - \vec{b}^\top\vec{x} + c \\
&\qquad \text{by adding and subtracting the same terms} \\
&= \frac{1}{2}(\vec{x} - \vec{x}^*)^\top A(\vec{x} - \vec{x}^*) + \vec{x}^\top\vec{b} - \frac{1}{2}(\vec{x}^*)^\top\vec{b} - \vec{b}^\top\vec{x} + c \text{ since } A\vec{x}^* = \vec{b} \\
&= \frac{1}{2}(\vec{x} - \vec{x}^*)^\top A(\vec{x} - \vec{x}^*) + \text{const. since the } \vec{x}^\top\vec{b} \text{ terms cancel}
\end{aligned}$$

Thus, up to a constant shift f is the same as the product $\frac{1}{2}(\vec{x} - \vec{x}^*)^\top A(\vec{x} - \vec{x}^*)$. In practice, we do not know $\vec{x}^*$, but this observation shows us the nature of f: It measures the distance from $\vec{x}$ to $\vec{x}^*$ with respect to the "A-norm" $\|\vec{v}\|_A^2 \equiv \vec{v}^\top A\vec{v}$.

Since A is symmetric and positive definite, even if it might be slow to carry out algorithmically, we know from §3.2.1 that A admits a Cholesky factorization $A = LL^\top$. With this factorization, f takes a nicer form:

$$f(\vec{x}) = \frac{1}{2}\|L^\top(\vec{x} - \vec{x}^*)\|_2^2 + \text{const.}$$

From this form of $f(\vec{x})$, we now know that the A-norm truly measures a distance between $\vec{x}$ and $\vec{x}^*$.

Define $\vec{y} \equiv L^\top\vec{x}$ and $\vec{y}^* \equiv L^\top\vec{x}^*$. After this change of variables, we are minimizing $\bar{f}(\vec{y}) \equiv \|\vec{y} - \vec{y}^*\|_2^2$. Optimizing $\bar{f}$ would be easy if we knew L and $\vec{y}^*$ (take $\vec{y} = \vec{y}^*$), but to eventually remove the need for L we consider the possibility of minimizing $\bar{f}$ using only line searches derived in §10.1.1.

We make an observation about minimizing our simplified function $\bar{f}$ using line searches, illustrated in Figure 10.3:

Proposition 10.1. Suppose $\{\vec{w}_1, \ldots, \vec{w}_n\}$ are orthogonal in $\mathbb{R}^n$. Then, $\bar{f}$ is minimized in at most n steps by line searching in direction $\vec{w}_1$, then direction $\vec{w}_2$, and so on.

Proof. Take the columns of $Q \in \mathbb{R}^{n \times n}$ to be the vectors $\vec{w}_i$; Q is an orthogonal matrix. Since Q is orthogonal, we can write $\bar{f}(\vec{y}) = \|\vec{y} - \vec{y}^*\|_2^2 = \|Q^\top \vec{y} - Q^\top \vec{y}^*\|_2^2$; in other words, we rotate so that $\vec{w}_1$ is the first standard basis vector, $\vec{w}_2$ is the second, and so on. If we write $\vec{z} \equiv Q^\top \vec{y}$ and $\vec{z}^* \equiv Q^\top \vec{y}^*$, then after the first iteration we must have $z_1 = z_1^*$, after the second iteration $z_2 = z_2^*$, and so on. After n steps we reach $z_n = z_n^*$, yielding the desired result. □

So, optimizing $\bar{f}$ can be accomplished via n line searches so long as those searches are in *orthogonal* directions.

All we did to pass from f to $\bar{f}$ is change coordinates using $L^\top$. Such a linear transformation takes straight lines to straight lines, so doing a line search on $\bar{f}$ along some vector $\vec{w}$ is equivalent to doing a line search along $(L^\top)^{-1}\vec{w}$ on our original quadratic function f. Conversely, if we do n line searches on f on directions $\vec{v}_i$ such that $L^\top \vec{v}_i \equiv \vec{w}_i$ are orthogonal, then by Proposition 10.1 we must have found $\vec{x}^*$. Requiring $\vec{w}_i \cdot \vec{w}_j = 0$ is the same as requiring

$$0 = \vec{w}_i \cdot \vec{w}_j = (L^\top \vec{v}_i)^\top (L^\top \vec{v}_j) = \vec{v}_i^\top (LL^\top)\vec{v}_j = \vec{v}_i^\top A \vec{v}_j.$$

We have just argued a corollary to Proposition 10.1. Define *conjugate* vectors as follows:

Definition 10.1 (A-conjugate vectors)**.** Two vectors $\vec{v}, \vec{w}$ are *A-conjugate* if $\vec{v}^\top A \vec{w} = 0$.

Then, based on our discussion we have shown:

Proposition 10.2. Suppose $\{\vec{v}_1, \ldots, \vec{v}_n\}$ are A-conjugate. Then, f is minimized in at most n steps by line searching in direction $\vec{v}_1$, then direction $\vec{v}_2$, and so on.

Inspired by this proposition, the conjugate gradients algorithm generates and searches along A-conjugate directions rather than moving along $-\nabla f$. This result might appear somewhat counterintuitive: We do *not* necessarily move along the steepest descent direction, but rather ask that our *set* of search directions satisfies a global criterion to make sure we do not repeat work. This setup guarantees convergence in a finite number of iterations and acknowledges the structure of f in terms of $\bar{f}$ discussed above.

We motivated A-conjugate directions by their orthogonality after applying $L^\top$ from the factorization $A = LL^\top$. From this standpoint, we are dealing with two dot products: $\vec{x}_i \cdot \vec{x}_j$ and $\vec{y}_i \cdot \vec{y}_j \equiv (L^\top \vec{x}_i) \cdot (L^\top \vec{x}_j) = x_i^\top LL^\top \vec{x}_j = \vec{x}_i^\top A \vec{x}_j$. These two products will figure into our subsequent discussion in equal amounts, so we denote the "A-inner product" as

$$\langle \vec{u}, \vec{v} \rangle_A \equiv (L^\top \vec{u}) \cdot (L^\top \vec{v}) = \vec{u}^\top A \vec{v}.$$

10.2.2 Suboptimality of Gradient Descent

So far, we know that if we can find n A-conjugate search directions, we can solve $A\vec{x} = \vec{b}$ in n steps via line searches along these directions. What remains is to uncover a formula for finding these directions efficiently. To do so, we will examine one more property of gradient descent that will inspire a more refined algorithm.

Suppose we are at $\vec{x}_k$ during an iterative line search method on $f(\vec{x})$; we will call the direction of steepest descent of f at $\vec{x}_k$ the *residual* $\vec{r}_k \equiv \vec{b} - A\vec{x}_k$. We may not decide to do a line search along $\vec{r}_k$ as in gradient descent, since the gradient directions are not

necessarily A-conjugate. So, generalizing slightly, we will find $\vec{x}_{k+1}$ via line search along a yet-undetermined direction $\vec{v}_{k+1}$.

From our derivation of gradient descent in §10.1.1, we should choose $\vec{x}_{k+1} = \vec{x}_k + \alpha_{k+1}\vec{v}_{k+1}$, where

$$\alpha_{k+1} = \frac{\vec{v}_{k+1}^\top \vec{r}_k}{\vec{v}_{k+1}^\top A\vec{v}_{k+1}}.$$

Applying this expansion of $\vec{x}_{k+1}$, we can write an update formula for the residual:

$$\begin{aligned}
\vec{r}_{k+1} &= \vec{b} - A\vec{x}_{k+1} \\
&= \vec{b} - A(\vec{x}_k + \alpha_{k+1}\vec{v}_{k+1}) \text{ by definition of } \vec{x}_{k+1} \\
&= (\vec{b} - A\vec{x}_k) - \alpha_{k+1}A\vec{v}_{k+1} \\
&= \vec{r}_k - \alpha_{k+1}A\vec{v}_{k+1} \text{ by definition of } \vec{r}_k
\end{aligned}$$

This formula holds regardless of our choice of $\vec{v}_{k+1}$ and can be applied to any iterative line search method.

In the case of gradient descent, we chose $\vec{v}_{k+1} \equiv \vec{r}_k$, giving a recurrence relation $\vec{r}_{k+1} = \vec{r}_k - \alpha_{k+1}A\vec{r}_k$. This formula inspires an instructive proposition:

Proposition 10.3. When performing gradient descent on f, $\operatorname{span}\{\vec{r}_0, \ldots, \vec{r}_k\} = \operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^k\vec{r}_0\}$.

Proof. This statement follows inductively from our formula for $\vec{r}_{k+1}$ above. □

The structure we are uncovering is beginning to look a lot like the Krylov subspace methods mentioned in Chapter 5: This is not a coincidence!

Gradient descent gets to $\vec{x}_k$ by moving along $\vec{r}_0$, then $\vec{r}_1$, and so on through $\vec{r}_k$. In the end we know that the iterate $\vec{x}_k$ of gradient descent on f lies somewhere in the plane $\vec{x}_0 + \operatorname{span}\{\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_{k-1}\} = \vec{x}_0 + \operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}$, by Proposition 10.3. Unfortunately, it is *not* true that if we run gradient descent, the iterate $\vec{x}_k$ is optimal in this subspace. In other words, it can be the case that

$$\vec{x}_k - \vec{x}_0 \neq \operatorname*{arg\,min}_{\vec{v} \in \operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}} f(\vec{x}_0 + \vec{v})$$

Ideally, switching this inequality to an equality would make sure that generating $\vec{x}_{k+1}$ from $\vec{x}_k$ does not "cancel out" any work done during iterations 1 to $k-1$.

If we reexamine our proof of Proposition 10.1 from this perspective, we can make an observation suggesting how we might use conjugacy to improve gradient descent. In particular, once z_i switches to z_i^*, it never changes value in a future iteration. After rotating back from $\vec{z}$ to $\vec{x}$ the following proposition holds:

Proposition 10.4. Take $\vec{x}_k$ to be the k-th iterate of the process from Proposition 10.1 after searching along $\vec{v}_k$. Then,

$$\vec{x}_k - \vec{x}_0 = \operatorname*{arg\,min}_{\vec{v} \in \operatorname{span}\{\vec{v}_1, \ldots, \vec{v}_k\}} f(\vec{x}_0 + \vec{v})$$

In the best of all possible worlds and in an attempt to outdo gradient descent, we might hope to find A-conjugate directions $\{\vec{v}_1, \ldots, \vec{v}_n\}$ such that $\operatorname{span}\{\vec{v}_1, \ldots, \vec{v}_k\} = \operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}$ for each k. Then by the previous two propositions, the resulting iterative scheme is guaranteed to do no worse than gradient descent even if it is halted early. But, we wish to do so without incurring significant memory demand or computation time. Amazingly, the conjugate gradient algorithm satisfies all these criteria.

10.2.3 Generating A-Conjugate Directions

Given any set of directions spanning $\mathbb{R}^n$, we can make them A-orthogonal using a method like Gram-Schmidt orthogonalization. Explicitly orthogonalizing $\{\vec{r}_0, A\vec{r}_0, A^2\vec{r}_0, \ldots\}$ to find the set of search directions is expensive and would require us to maintain a complete list of directions in memory; this construction likely would exceed the time and memory requirements even of Gaussian elimination. Alternatively, we will reveal one final observation *about* Gram-Schmidt that makes conjugate gradients tractable by generating conjugate directions without an expensive orthogonalization process.

Ignoring these issues, we might write a "method of conjugate directions" using the following iterations:

$$\begin{aligned}
&\vec{v}_k \leftarrow A^{k-1}\vec{r}_0 - \textstyle\sum_{i<k} \frac{\langle A^{k-1}\vec{r}_0, \vec{v}_i\rangle_A}{\langle \vec{v}_i, \vec{v}_i\rangle_A}\vec{v}_i && \triangleright \text{ Explicit Gram-Schmidt}\\
&\alpha_k \leftarrow \frac{\vec{v}_k^\top \vec{r}_{k-1}}{\vec{v}_k^\top A\vec{v}_k} && \triangleright \text{ Line search}\\
&\vec{x}_k \leftarrow \vec{x}_{k-1} + \alpha_k\vec{v}_k && \triangleright \text{ Update estimate}\\
&\vec{r}_k \leftarrow \vec{r}_{k-1} - \alpha_k A\vec{v}_k && \triangleright \text{ Update residual}
\end{aligned}$$

Here, we compute the k-th search direction $\vec{v}_k$ by projecting $\vec{v}_1, \ldots, \vec{v}_{k-1}$ out of the vector $A^{k-1}\vec{r}_0$. This algorithm has the property $\text{span}\{\vec{v}_1, \ldots, \vec{v}_k\} = \text{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}$ suggested in §10.2.2, but it has two issues:

1. Similar to power iteration for eigenvectors, the power $A^{k-1}\vec{r}_0$ is likely to look mostly like the first eigenvector of A, making projection poorly conditioned when k is large.
2. We have to store $\vec{v}_1, \ldots, \vec{v}_{k-1}$ to compute $\vec{v}_k$, so each iteration needs more memory and time than the last.

We can fix the first issue in a relatively straightforward manner. Right now, we project the previous search directions out of $A^{k-1}\vec{r}_0$, but in reality we can project out previous directions from *any* vector $\vec{w}$ so long as

$$\vec{w} \in \text{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\} \backslash \text{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-2}\vec{r}_0\},$$

that is, as long as $\vec{w}$ has some component in the new part of the space.

An alternative choice of $\vec{w}$ in this span is the residual $\vec{r}_{k-1}$. We can check this using the residual update $\vec{r}_k = \vec{r}_{k-1} - \alpha_k A\vec{v}_k$; in this expression, we multiply $\vec{v}_k$ by A, introducing the new power of A that we need. This choice also more closely mimics the gradient descent algorithm, which took $\vec{v}_k = \vec{r}_{k-1}$. We can update our algorithm to use this improved choice:

$$\begin{aligned}
&\vec{v}_k \leftarrow \vec{r}_{k-1} - \textstyle\sum_{i<k} \frac{\langle \vec{r}_{k-1}, \vec{v}_i\rangle_A}{\langle \vec{v}_i, \vec{v}_i\rangle_A}\vec{v}_i && \triangleright \text{ Gram-Schmidt on residual}\\
&\alpha_k \leftarrow \frac{\vec{v}_k^\top \vec{r}_{k-1}}{\vec{v}_k^\top A\vec{v}_k} && \triangleright \text{ Line search}\\
&\vec{x}_k \leftarrow \vec{x}_{k-1} + \alpha_k\vec{v}_k && \triangleright \text{ Update estimate}\\
&\vec{r}_k \leftarrow \vec{r}_{k-1} - \alpha_k A\vec{v}_k && \triangleright \text{ Update residual}
\end{aligned}$$

Now we do not do arithmetic with the poorly-conditioned vector $A^{k-1}\vec{r}_0$ but still have the "memory" problem above since the sum in the first step is over $k-1$ vectors.

A surprising observation about the residual Gram-Schmidt step above is that most terms in the sum are exactly zero! This amazing observation allows each iteration of conjugate gradients to be carried out without increasing memory requirements. We memorialize this result in a proposition:

Proposition 10.5. In the second "conjugate direction" method above, $\langle \vec{r}_k, \vec{v}_\ell \rangle_A = 0$ for all $\ell < k$.

Proof. We proceed inductively. There is nothing to prove for the base case $k = 1$, so assume $k > 1$ and that the result holds for all $k' < k$. By the residual update formula, we know:

$$\langle \vec{r}_k, \vec{v}_\ell \rangle_A = \langle \vec{r}_{k-1}, \vec{v}_\ell \rangle_A - \alpha_k \langle A\vec{v}_k, \vec{v}_\ell \rangle_A = \langle \vec{r}_{k-1}, \vec{v}_\ell \rangle_A - \alpha_k \langle \vec{v}_k, A\vec{v}_\ell \rangle_A,$$

where the second equality follows from symmetry of A.

First, suppose $\ell < k-1$. Then the first term of the difference above is zero by induction. Furthermore, by construction $A\vec{v}_\ell \in \operatorname{span}\{\vec{v}_1, \ldots, \vec{v}_{\ell+1}\}$, so since we have constructed our search directions to be A-conjugate we know the second term must be zero as well.

To conclude the proof, we consider the case $\ell = k-1$. Using the residual update formula, we know:

$$A\vec{v}_{k-1} = \frac{1}{\alpha_{k-1}}(\vec{r}_{k-2} - \vec{r}_{k-1})$$

Premultiplying by $\vec{r}_k$ shows:

$$\langle \vec{r}_k, \vec{v}_{k-1} \rangle_A = \frac{1}{\alpha_{k-1}} \vec{r}_k^\top (\vec{r}_{k-2} - \vec{r}_{k-1})$$

The difference $\vec{r}_{k-2} - \vec{r}_{k-1}$ lives in $\operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}$, by the residual update formula. Proposition 10.4 shows that $\vec{x}_k$ is optimal in this subspace. Since $\vec{r}_k = -\nabla f(\vec{x}_k)$, this implies that we must have $\vec{r}_k \perp \operatorname{span}\{\vec{r}_0, A\vec{r}_0, \ldots, A^{k-1}\vec{r}_0\}$, since otherwise there would exist a direction in the subspace to move from $\vec{x}_k$ to decrease f. In particular, this shows the inner product above $\langle \vec{r}_k, \vec{v}_{k-1} \rangle_A = 0$, as desired. □

Thus, our proof above shows that we can find a new direction $\vec{v}_k$ as follows:

$$\begin{aligned}
\vec{v}_k &= \vec{r}_{k-1} - \sum_{i<k} \frac{\langle \vec{r}_{k-1}, \vec{v}_i \rangle_A}{\langle \vec{v}_i, \vec{v}_i \rangle_A} \vec{v}_i \text{ by the Gram-Schmidt formula} \\
&= \vec{r}_{k-1} - \frac{\langle \vec{r}_{k-1}, \vec{v}_{k-1} \rangle_A}{\langle \vec{v}_{k-1}, \vec{v}_{k-1} \rangle_A} \vec{v}_{k-1} \text{ because the remaining terms vanish}
\end{aligned}$$

Since the summation over i disappears, the cost of computing $\vec{v}_k$ has no dependence on k.

10.2.4 Formulating the Conjugate Gradients Algorithm

Now that we can obtain A-conjugate search directions with relatively little computational effort, we apply this strategy to formulate the conjugate gradients algorithm, with full pseudocode in Figure 10.4(a):

$$\begin{array}{ll}
\vec{v}_k \leftarrow \vec{r}_{k-1} - \frac{\langle \vec{r}_{k-1}, \vec{v}_{k-1} \rangle_A}{\langle \vec{v}_{k-1}, \vec{v}_{k-1} \rangle_A} \vec{v}_{k-1} & \triangleright \text{ Update search direction} \\
\alpha_k \leftarrow \frac{\vec{v}_k^\top \vec{r}_{k-1}}{\vec{v}_k^\top A\vec{v}_k} & \triangleright \text{ Line search} \\
\vec{x}_k \leftarrow \vec{x}_{k-1} + \alpha_k \vec{v}_k & \triangleright \text{ Update estimate} \\
\vec{r}_k \leftarrow \vec{r}_{k-1} - \alpha_k A\vec{v}_k & \triangleright \text{ Update residual}
\end{array}$$

This iterative scheme is only a minor adjustment to the gradient descent algorithm but has many desirable properties by construction:

- $f(\vec{x}_k)$ is upper-bounded by that of the k-th iterate of gradient descent.

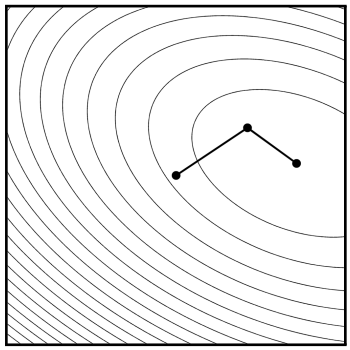
function CONJUGATE-GRAD-1($A, \vec{b}, \vec{x}_0$)
  $\vec{x} \leftarrow \vec{x}_0$
  $\vec{r} \leftarrow \vec{b} - A\vec{x}$
  $\vec{v} \leftarrow \vec{r}$
  for $k \leftarrow 1, 2, 3, \ldots$
    $\alpha \leftarrow \frac{\vec{v}^\top \vec{r}}{\vec{v}^\top A\vec{v}}$ ▷ Line search
    $\vec{x} \leftarrow \vec{x} + \alpha\vec{v}$ ▷ Update estimate
    $\vec{r} \leftarrow \vec{r} - \alpha A\vec{v}$ ▷ Update residual
    if $\|\vec{r}\|_2^2 < \varepsilon\|\vec{r}_0\|_2^2$ **then**
      return $x^* = \vec{x}$
    $\vec{v} \leftarrow \vec{r} - \frac{\langle \vec{r}, \vec{v} \rangle_A}{\langle \vec{v}, \vec{v} \rangle_A}\vec{v}$ ▷ Search direction

(a)

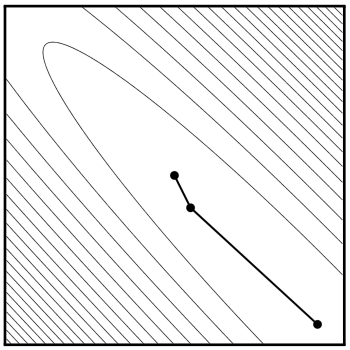
function CONJUGATE-GRAD-2($A, \vec{b}, \vec{x}_0$)
  $\vec{x} \leftarrow \vec{x}_0$
  $\vec{r} \leftarrow \vec{b} - A\vec{x}$
  $\vec{v} \leftarrow \vec{r}$
  $\beta \leftarrow 0$
  for $k \leftarrow 1, 2, 3, \ldots$
    $\vec{v} \leftarrow \vec{r} + \beta\vec{v}$ ▷ Search direction
    $\alpha \leftarrow \frac{\|\vec{r}\|_2^2}{\vec{v}^\top A\vec{v}}$ ▷ Line search
    $\vec{x} \leftarrow \vec{x} + \alpha\vec{v}$ ▷ Update estimate
    $\vec{r}_{\text{old}} \leftarrow \vec{r}$ ▷ Save old residual
    $\vec{r} \leftarrow \vec{r} - \alpha A\vec{v}$ ▷ Update residual
    if $\|\vec{r}\|_2^2 < \varepsilon\|\vec{r}_0\|_2^2$ **then**
      return $x^* = \vec{x}$
    $\beta \leftarrow \|\vec{r}\|_2^2 / \|\vec{r}_{\text{old}}\|_2^2$ ▷ Direction step

(b)

FIGURE 10.4 Two equivalent formulations of the conjugate gradients algorithm for solving $A\vec{x} = \vec{b}$ when A is symmetric and positive definite. The initial guess $\vec{x}_0$ can be $\vec{0}$ in the absence of a better estimate.



Well-conditioned A



Poorly-conditioned A

FIGURE 10.5 The conjugate gradients algorithm solves both linear systems in Figure 10.2 in two steps.

- The algorithm converges to $\vec{x}^*$ in n steps, as illustrated in Figure 10.5.
- At each step, the iterate $\vec{x}_k$ is optimal in the subspace spanned by the first k search directions.

In the interests of squeezing maximal numerical quality out of conjugate gradients, we can try to simplify the numerics of the expressions in our formulation in Figure 10.4(a). For instance, if we plug the search direction update into the formula for α_k, by orthogonality we can write:

$$\alpha_k = \frac{\vec{r}_{k-1}^\top \vec{r}_{k-1}}{\vec{v}_k^\top A \vec{v}_k}$$

The numerator of this fraction now is guaranteed to be nonnegative without numerical precision issues.

Similarly, we can define a constant β_k to split the search direction update into two steps:

$$\beta_k \equiv -\frac{\langle \vec{r}_{k-1}, \vec{v}_{k-1} \rangle_A}{\langle \vec{v}_{k-1}, \vec{v}_{k-1} \rangle_A}$$
$$\vec{v}_k = \vec{r}_{k-1} + \beta_k \vec{v}_{k-1}$$

We can simplify the formula for β_k:

$$\begin{aligned}
\beta_k &= -\frac{\vec{r}_{k-1} A \vec{v}_{k-1}}{\vec{v}_{k-1}^\top A \vec{v}_{k-1}} \text{ by definition of } \langle \cdot, \cdot \rangle_A \\
&= -\frac{\vec{r}_{k-1}^\top (\vec{r}_{k-2} - \vec{r}_{k-1})}{\alpha_{k-1} \vec{v}_{k-1}^\top A \vec{v}_{k-1}} \text{ since } \vec{r}_k = \vec{r}_{k-1} - \alpha_k A \vec{v}_k \\
&= \frac{\vec{r}_{k-1}^\top \vec{r}_{k-1}}{\alpha_{k-1} \vec{v}_{k-1}^\top A \vec{v}_{k-1}} \text{ by a calculation below} \\
&= \frac{\vec{r}_{k-1}^\top \vec{r}_{k-1}}{\vec{r}_{k-2}^\top \vec{r}_{k-2}} \text{ by our last formula for } \alpha_k
\end{aligned}$$

This expression reveals that $\beta_k \geq 0$, a property which might not have held after numerical precision issues. We have one remaining calculation below:

$$\begin{aligned}
\vec{r}_{k-2}^\top \vec{r}_{k-1} &= \vec{r}_{k-2}^\top (\vec{r}_{k-2} - \alpha_{k-1} A \vec{v}_{k-1}) \text{ by our residual update formula} \\
&= \vec{r}_{k-2}^\top \vec{r}_{k-2} - \frac{\vec{r}_{k-2}^\top \vec{r}_{k-2}}{\vec{v}_{k-1}^\top A \vec{v}_{k-1}} \vec{r}_{k-2}^\top A \vec{v}_{k-1} \text{ by our formula for } \alpha_k \\
&= \vec{r}_{k-2}^\top \vec{r}_{k-2} - \frac{\vec{r}_{k-2}^\top \vec{r}_{k-2}}{\vec{v}_{k-1}^\top A \vec{v}_{k-1}} \vec{v}_{k-1}^\top A \vec{v}_{k-1} \text{ by the update for } \vec{v}_k \text{ and } A\text{-conjugacy of the } \vec{v}_k\text{'s} \\
&= 0, \text{ as needed.}
\end{aligned}$$

Our new observations about the iterates of CG provide an alternative but equivalent formulation, shown in Figure 10.4(b), that can have better numerical properties. Also for numerical reasons, occasionally rather than using the update formula for $\vec{r}_k$ it is advisable to use the residual formula $\vec{r}_k = \vec{b} - A\vec{x}_k$. This formula requires an extra matrix-vector multiply but repairs numerical "drift" caused by finite-precision rounding. There is no need to store a long list of previous residuals or search directions; conjugate gradients takes a constant amount of space from iteration to iteration.

10.2.5 Convergence and Stopping Conditions

By construction, the conjugate gradients (CG) algorithm is guaranteed to converge as fast as gradient descent on f, while being no harder to implement and having a number of other favorable properties. A detailed discussion of CG convergence is out of the scope of our discussion, but in general the algorithm behaves best on matrices with evenly-distributed eigenvalues over a small range. One rough estimate paralleling our estimate in §10.1.2 shows that the CG algorithm satisfies:

$$\frac{f(\vec{x}_k) - f(\vec{x}^*)}{f(\vec{x}_0) - f(\vec{x}^*)} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$$

where $\kappa \equiv \operatorname{cond} A$. More generally, the number of iterations needed for conjugate gradient to reach a given error value usually can be bounded by a function of $\sqrt{\kappa}$, whereas bounds for convergence of gradient descent are proportional to κ.

We know that conjugate gradients is guaranteed to converge to $\vec{x}^*$ exactly in n steps—m steps if A has $m < n$ unique eigenvalues—but when n is large it may be preferable to stop earlier than that. The formula for β_k will divide by zero when the residual gets very short, which can cause numerical precision issues near the minimum of f. Thus, in practice CG usually his halted when the ratio $\|\vec{r}_k\|/\|\vec{r}_0\|$ is sufficiently small.

10.3 PRECONDITIONING

We now have two powerful iterative schemes for finding solutions to $A\vec{x} = \vec{b}$ when A is symmetric and positive definite, gradient descent and conjugate gradients. Each converges *unconditionally*, meaning that regardless of the initial guess $\vec{x}_0$ with enough iterations they will get arbitrarily close to the true solution $\vec{x}^*$; conjugate gradients even reaches $\vec{x}^*$ exactly in a finite number of iterations. The "clock time" taken to solve $A\vec{x} = \vec{b}$ for both of these methods is proportional to the number of iterations needed to reach $\vec{x}^*$ within an acceptable tolerance, so it makes sense to minimize the number of iterations until convergence.

We are able to characterize the convergence rates of both algorithms and other related iterative techniques in terms of the condition number $\operatorname{cond} A$. That is, the smaller the value of $\operatorname{cond} A$, the less time it should take to solve $A\vec{x} = \vec{b}$. This situation contrasts with Gaussian elimination, which takes the same number of steps regardless of A; what is new here is that the conditioning of A affects not only the quality of the output of iterative methods but also the speed at which $\vec{x}^*$ is approached.

For any invertible matrix P, solving $PA\vec{x} = P\vec{b}$ is equivalent to solving $A\vec{x} = \vec{b}$. The condition number of PA, however, does *not* need to be the same as that of A. In the extreme, if we took $P = A^{-1}$ then conditioning issues would be removed altogether! More generally, suppose $P \approx A^{-1}$. Then, we expect $\operatorname{cond} PA \ll \operatorname{cond} A$, making it advisable to apply P before solving the linear system using iterative methods. In this case, we will call P a *preconditioner*.

While the idea of preconditioning appears attractive, two issues remain:

1. While A may be symmetric and positive definite, the product PA in general will not enjoy these properties.
2. We need to find $P \approx A^{-1}$ that is easier to compute than A^{-1} itself.

We address these issues in the sections below.

10.3.1 CG with Preconditioning

We will focus our discussion of preconditioning on conjugate gradients since it has better convergence properties than gradient descent, although most of our constructions can be paralleled to precondition other iterative linear methods.

Reviewing the initial steps in §10.2.1, we see that our construction of CG fundamentally depends on both the symmetry and positive definiteness of A. Hence, running CG on PA usually will not converge out of the box. Suppose, however, that the preconditioner P is itself symmetric and positive definite. This is a reasonable assumption since A^{-1} must satisfy these properties. Then, we can write a Cholesky factorization of the inverse $P^{-1} = EE^\top$. We make the following observation:

Proposition 10.6. The condition number of PA is the same as that of $E^{-1}AE^{-\top}$.

Proof. We show that PA and $E^{-1}AE^{-\top}$ have the same singular values; the condition number is the ratio of the maximum singular value to the minimum singular value, so this fact is more than sufficient to prove the proposition. Since E is invertible and A is symmetric and positive definite, $E^{-1}AE^{-\top}$ must also be symmetric and positive definite. For this reason, the eigenvalues of $E^{-1}AE^{-\top}$ are its singular values. Suppose $E^{-1}AE^{-\top}\vec{x} = \lambda\vec{x}$. By construction, $P^{-1} = EE^\top$, so $P = E^{-\top}E^{-1}$., Thus, if we pre-multiply both sides of our eigenvector expression by $E^{-\top}$ we find $PAE^{-\top}\vec{x} = \lambda E^{-\top}\vec{x}$. Defining $\vec{y} \equiv E^{-\top}\vec{x}$ shows $PA\vec{y} = \lambda\vec{y}$. Each of these steps is reversible, showing that PA and $E^{-1}AE^{-\top}$ both have full eigenspaces and identical eigenvalues. □

This proposition implies that if we do CG on the symmetric positive definite matrix $E^{-1}AE^{-\top}$, we will receive the same conditioning benefits enjoyed by PA. As in the proof of Proposition 10.6, we could carry out our new solve for $\vec{y} = E^\top\vec{x}$ in two steps:

1. Solve $E^{-1}AE^{-\top}\vec{y} = E^{-1}\vec{b}$ for $\vec{y}$.
2. Multiply to find $\vec{x} = E^{-\top}\vec{y}$.

Finding E or its inverse would be integral to this strategy, but doing so can induce fill and take too much time. By modifying the steps of CG for the first step above, however, we can make this factorization unnecessary.

If we had computed E, we could perform step 1 using CG as follows:

$$
\begin{array}{ll}
\beta_k \leftarrow \frac{\vec{r}_{k-1}^\top\vec{r}_{k-1}}{\vec{r}_{k-2}^\top\vec{r}_{k-2}} & \triangleright \text{ Update search direction} \\
\vec{v}_k \leftarrow \vec{r}_{k-1} + \beta_k\vec{v}_{k-1} & \\
\alpha_k \leftarrow \frac{\vec{r}_{k-1}^\top\vec{r}_{k-1}}{\vec{v}_k^\top E^{-1}AE^{-\top}\vec{v}_k} & \triangleright \text{ Line search} \\
\vec{y}_k \leftarrow \vec{y}_{k-1} + \alpha_k\vec{v}_k & \triangleright \text{ Update estimate} \\
\vec{r}_k \leftarrow \vec{r}_{k-1} - \alpha_k E^{-1}AE^{-\top}\vec{v}_k & \triangleright \text{ Update residual}
\end{array}
$$

This iterative scheme will converge according to the conditioning of $E^{-1}AE^{-\top}$.

Define $\tilde{r}_k \equiv E\vec{r}_k$, $\tilde{v}_k \equiv E^{-\top}\vec{v}_k$, and $\vec{x}_k \equiv E\vec{y}_k$. By the relationship $P = E^{-\top}E^{-1}$, we can rewrite our preconditioned conjugate gradients iteration completely in terms of these new variables:

$$
\begin{aligned}
\beta_k &\leftarrow \frac{\tilde{r}_{k-1}^\top P \tilde{r}_{k-1}}{\tilde{r}_{k-2}^\top P \tilde{r}_{k-2}} && \triangleright \text{ Update search direction} \\
\tilde{v}_k &\leftarrow P\tilde{r}_{k-1} + \beta_k \tilde{v}_{k-1} \\
\alpha_k &\leftarrow \frac{\vec{r}_{k-1}^\top P \vec{r}_{k-1}}{\tilde{v}_k^\top A \tilde{v}_k} && \triangleright \text{ Line search} \\
\vec{x}_k &\leftarrow \vec{x}_{k-1} + \alpha_k \tilde{v}_k && \triangleright \text{ Update estimate} \\
\tilde{r}_k &\leftarrow \tilde{r}_{k-1} - \alpha_k A \tilde{v}_k && \triangleright \text{ Update residual}
\end{aligned}
$$

This iteration does not depend on the Cholesky factorization of P^{-1}, but instead can be carried out using only P and A. By the substitutions above, $\vec{x}_k \to \vec{x}^*$, and this scheme enjoys the benefits of preconditioning without needing to factor the preconditioner.

As a side note, more general preconditioning can be carried out by replacing A with PAQ for a second matrix Q, although this second matrix will require additional computations to apply. This extension represents a common trade-off: If a preconditioner takes too long to apply in each iteration of CG, it may not be worth the reduced number of iterations.

10.3.2 Common Preconditioners

Finding good preconditioners in practice is as much an art as it is a science. Finding an effective approximation P of A^{-1} depends on the structure of A, the particular application at hand, and so on. Even rough approximations, however, can help convergence, so rarely do applications of CG appear that do *not* use a preconditioner.

The best strategy for finding P often is application-specific, and generally it is necessary to test a few possibilities for P before settling on the most effective option. A few common generic preconditioners include the following:

- A *diagonal* (or "*Jacobi*") preconditioner takes P to be the matrix obtained by inverting diagonal elements of A; that is, P is the diagonal matrix with entries $1/a_{ii}$. This preconditioner can alleviate nonuniform scaling from row to row, which is a common cause of poor conditioning.

- The *sparse approximate inverse* preconditioner is formulated by solving a subproblem $\min_{P \in S} \|AP - I\|_{\text{Fro}}$, where P is restricted to be in a set S of matrices over which it is less difficult to optimize such an objective. For instance, a common constraint is to prescribe a sparsity pattern for P, e.g. that it only has nonzeros on its diagonal or where A has nonzeros.

- The *incomplete Cholesky* preconditioner factors $A \approx L_* L_*^\top$ and then approximates A^{-1} by solving the appropriate forward- and back-substitution problems. For instance, a popular heuristic involves going through the steps of Cholesky factorization but only saving the output in positions (i, j) where $a_{ij} \neq 0$.

- The nonzero values in A can be used to construct a graph with edge (i, j) whenever $a_{ij} \neq 0$. Removing edges in the graph or grouping nodes may disconnect assorted components; the resulting system is block-diagonal after permuting rows and columns and thus can be solved using a sequence of smaller solves. Such a *domain decomposition* can be effective for linear systems arising from differential equations such as those considered in Chapter 15.

Some preconditioners come with bounds describing changes to the conditioning of A after replacing it with PA, but for the most part these are heuristic strategies that should be tested and refined.

10.4 OTHER ITERATIVE SCHEMES

The algorithms we have developed in detail this chapter apply for solving $A\vec{x} = \vec{b}$ when A is square, symmetric, and positive definite. We have focused on this case because it appears so often in practice, but there are cases when A is asymmetric, indefinite, or even rectangular. It is out of the scope of our discussion to derive iterative algorithms in each case, since many require some specialized analysis or advanced development (see e.g. [7, 46, 52, 100]), but we summarize some techniques here from a high-level:

- *Splitting* methods decompose $A = M - N$ and use the fact that $A\vec{x} = \vec{b}$ is equivalent to $M\vec{x} = N\vec{x} + \vec{b}$. If M is easy to invert, then a fixed-point scheme can be derived by writing $M\vec{x}_k = N\vec{x}_{k-1} + \vec{b}$; these techniques are easy to implement but have convergence depending on the spectrum of the matrix $G = M^{-1}N$ and in particular can diverge when the spectral radius of G is greater than one. One popular choice of M is the diagonal of A. Methods such as *successive over-relaxation* (SOR) weight these two terms for better convergence.

- The *conjugate gradient normal equation residual* (CGNR) method applies the CG algorithm to the normal equations $A^\top A\vec{x} = A^\top\vec{b}$. This method is guaranteed to converge so long as A is full-rank, but convergence can be slow thanks to poor conditioning of $A^\top A$ as in §4.1.

- The *conjugate gradient normal equation error* (CGNE) method similarly solves $AA^\top\vec{y} = \vec{b}$; then the solution of $A\vec{x} = \vec{b}$ is $A^\top\vec{y}$.

- Methods such as MINRES and SYMMLQ apply to all symmetric matrices A by replacing our quadratic form $f(\vec{x})$ with $g(\vec{x}) \equiv \|\vec{b} - A\vec{x}\|_2$ [88]; this function g is minimized at solutions to $A\vec{x} = \vec{b}$ regardless of the definiteness of A.

- Given the poor conditioning of CGNR and CGNE, the LSQR and LSMR algorithms also minimize $g(\vec{x})$ with fewer assumptions on A, in particular allowing for solution of least-squares systems [89, 39].

- Generalized methods including GMRES, QMR, BiCG, CGS, and BiCGStab solve $A\vec{x} = \vec{b}$ with the only caveat that A is square and invertible [101, 41, 37, 109, 120]. They optimize similar energies but often have to store more information about previous iterations and may have to factor intermediate matrices to guarantee convergence with such generality.

- Finally, methods like the *Fletcher-Reeves*, *Hestenes-Stiefel*, *Polak-Ribière*, and *Dai-Yuan* algorithms return to the more general problem of minimizing a non-quadratic function f, applying conjugate gradient steps to finding new line search directions [28, 38, 55, 95]. Functions f that are well-approximated by quadratics can be minimized very effectively using these strategies, even though they do not necessarily make use of the Hessian. For instance, the Fletcher-Reeves method replaces the residual in CG iterations with the negative gradient $-\nabla f$.

Many of these algorithms are nearly as easy to implement as CG or gradient descent, and pre-packaged implementations are readily available that only require inputting A and $\vec{b}$. Many of the algorithms listed above require being able to multiply vectors both by A and by $A^\top$, which can be a technical challenge in some cases. As a rule of thumb, the more general a method is—that is, the fewer the assumptions a method makes on the structure of the matrix A—the more iterations it is likely to need to compensate for this lack of

assumptions. This said, there are no hard-and-fast rules that can be applied by examining the elements of A for guessing the most successful iterative scheme.

10.5 EXERCISES

10.1 If we use infinite-precision arithmetic (so rounding is not an issue), can the conjugate gradients algorithm be used to recover *exact* solutions to $A\vec{x} = \vec{b}$ for symmetric positive definite matrices A? Why or why not?

10.2 Suppose $A \in \mathbb{R}^{n\times n}$ is invertible but not symmetric or positive definite.

(a) Show that $A^\top A$ is symmetric and positive definite.

(b) Propose a strategy for solving $A\vec{x} = \vec{b}$ using the conjugate gradients algorithm based on your observation in (a).

(c) How quickly do you expect conjugate gradients to converge in this case? Why?

10.3 In this problem we will derive an iterative method of solving $A\vec{x} = \vec{b}$ via *splitting* [46].

(a) Suppose we decompose $A = M - N$, where M is invertible. Show that the iterative scheme $\vec{x}_k = M^{-1}(N\vec{x}_{k-1} + \vec{b})$ converges to $A^{-1}\vec{b}$ when $\max\{|\lambda| : \lambda \text{ is an eigenvalue of } M^{-1}N\} < 1$.
Hint: Define $\vec{x}^* = A^{-1}\vec{b}$ and take $\vec{e}_k = \vec{x}_k - \vec{x}^*$. Show that $\vec{e}_k = G^k\vec{e}_0$, where $G = M^{-1}N$. For this problem, you can assume that the eigenvectors of G span $\mathbb{R}^n$ (it is possible to prove this statement without the assumption but doing so requires more analysis than we have covered).

(b) Suppose A is strictly diagonally dominant, that is, for each i it satisfies

$$\sum_{j\neq i} |a_{ij}| < |a_{ii}|.$$

Suppose we define M to be the diagonal part of A and $N = M - A$. Show that the iterative scheme from part 10.3a converges in this case. You can assume the statement from 10.3a holds regardless of the eigenspace of G.

10.4 As introduced in §9.4.3, a graph is a data structure $G = (V, E)$ consisting of n vertices in a set $V = \{1, \ldots, n\}$ and a set of edges $E \subseteq V \times V$. A common problem is *graph layout*, where we choose positions of the vertices in V on the plane $\mathbb{R}^2$ respecting the connectivity of G. For this problem we will assume $(i, i) \notin E$ for all $i \in V$.

(a) Take $\vec{v}_1, \ldots, \vec{v}_n \in \mathbb{R}^2$ to be the positions of the vertices in V; these are the unknowns in graph layout. The Dirichlet energy of a layout is

$$E(\vec{v}_1, \ldots, \vec{v}_n) = \sum_{(i,j)\in E} \|\vec{v}_i - \vec{v}_j\|_2^2.$$

Suppose an artist specifies positions of vertices in a nonempty subset $V_0 \subseteq V$. We will label these positions as $\vec{v}_k^0$ for $k \in V_0$. Derive two $(n - |V_0|) \times (n - |V_0|)$

linear systems of equations satisfied by the x and y components of the unknown $\vec{v}_i$'s solving the following minimization problem:

$$\begin{aligned} &\text{minimize } E(\vec{v}_1, \ldots, \vec{v}_n) \\ &\text{such that } \vec{v}_k = \vec{v}_k^0 \ \forall k \in V_0 \end{aligned}$$

Hint: Your answer should yield two linear systems $A\vec{x} = \vec{b}_x$ and $A\vec{y} = \vec{b}_y$.

(b) Show that your systems from the previous part are symmetric and positive definite.

(c) Implement both gradient descent and conjugate gradients for solving this system, updating a display of the graph layout after each iteration. Compare the number of iterations needed to reach a reasonable solution using both strategies.

(d) Implement preconditioned conjugate gradients using a preconditioner of your choice. How much does convergence improve?

10.5[DH] The *successive over-relaxation* (SOR) method is an example of an iterative splitting method for solving $A\vec{x} = \vec{b}$. Suppose we decompose $A = D + L + U$, where D, L, and U are the diagonal, strictly lower triangular, and strictly upper triangular parts of A, respectively. Then, the SOR iteration is given by:

$$(\omega^{-1}D + L)\vec{x}_{k+1} = ((\omega^{-1} - 1)D - U)\vec{x}_k + \vec{b},$$

for some constant ω. We will show that if A is symmetric and positive definite and $\omega \in (0, 2)$, then the SOR method converges.

(a) Show how SOR is an instance of the splitting method in problem 10.3 by defining matrices M and N appropriately. Hence, using this problem we now only need to show that $\rho(G) < 1$ for $G = M^{-1}N$ to establish convergence of SOR.

(b) Define $Q \equiv (w^{-1}D + L)$ and let $\vec{y} = (I - G)\vec{x}$ for an arbitrary eigenvector $\vec{x} \in \mathbb{C}^n$ of G with corresponding eigenvalue $\lambda \in \mathbb{C}$. Derive expressions for $Q\vec{y}$ and $(Q - A)\vec{y}$ in terms of A, $\vec{x}$, and λ.

(c) Use the inner product $\langle \vec{x}, \vec{y} \rangle_A$ to conclude that $d_{ii} \equiv \langle \vec{e}_i, \vec{e}_i \rangle_A > 0$. This expression shows that all the possibly nonzero elements of the diagonal matrix D are positive.
Note: We are dealing with complex values here, so for the remainder of this problem inner products are given by $\langle \vec{x}, \vec{y} \rangle_A \equiv (A\vec{x})^\top \text{conjugate}(\vec{y})$.

(d) Substitute the definition of Q into your relationships from part 10.5b and simplify to show that:

$$\begin{aligned} \omega^{-1}\langle \vec{y}, \vec{y} \rangle_D + \langle \vec{y}, \vec{y} \rangle_L &= (1 - \bar{\lambda})\langle \vec{x}, \vec{x} \rangle_A \\ (\omega^{-1} - 1)\langle \vec{y}, \vec{y} \rangle_D - \langle \vec{y}, \vec{y} \rangle_U &= (1 - \lambda)\bar{\lambda}\langle \vec{x}, \vec{x} \rangle_A \end{aligned}$$

(e) Recalling our assumptions on A, what can you say about $\langle \vec{y}, \vec{y} \rangle_L$ and $\langle \vec{y}, \vec{y} \rangle_U$? Use this and the previous part to conclude that

$$(2\omega^{-1} - 1)\langle \vec{y}, \vec{y} \rangle_D = (1 - |\lambda|^2)\langle \vec{x}, \vec{x} \rangle_A.$$

(f) Justify why, under the given assumptions and results of the previous parts, each of $(2\omega^{-1}-1)$, $\langle \vec{y}, \vec{y} \rangle_D$, and $\langle \vec{x}, \vec{x} \rangle_A$ must be positive. What does this imply about $|\lambda|$? Conclude that the SOR method converges under our assumptions.

10.6[DH] ("Gradient domain painting," [81]) Let $I : S \to \mathbb{R}$ be a monochromatic image, where $S \subset \mathbb{R}^2$ is a rectangle. We know I on a collection of square pixels tiling S.

Suppose an artist is editing I in the gradient domain. This means the artist is edits the x and y derivatives g_x and g_y of I rather than values in I. After editing g_x and g_y, we need to recover a new image $\tilde{I}$ that has the edited gradients, at least approximately.

(a) For the artist to paint in the gradient domain, we first have to calculate discrete approximations of g_x and g_y using the values of I on different pixels. We will cover this problem in detail in Chapter 13. For now, how might you estimate the derivatives of I in the x and y directions from a pixel using the values of I at one or both of the two horizontally adjacent pixels?

(b) Describe matrices A_x and A_y such that $A_x I = g_x$ and $A_y I = g_y$, where in this case we have written I as a vector $I = [I_{1,1}, I_{1,n2}, ..., I_{1,n}, I_{2,1}, ..., I_{m,n}]^T$ and $I_{i,j}$ is the value of I at pixel (i, j). Assume the image I is m pixels tall and n pixels wide.

(c) Give an example of a function $g : \mathbb{R}^2 \to \mathbb{R}^2$ that is not a gradient, that is, g admits no f such that $\nabla f = g$. Justify your answer.

(d) In light of the fact that $\nabla \tilde{I} = g$ may not be solvable exactly, propose an optimization problem whose solution is the "best" approximate solution (in the L_2 norm) to this equation. Describe the advantage of using conjugate gradients to solve such a system.

10.7 The locally optimal block preconditioned conjugate gradient (LOBPCG) algorithm applies conjugate gradients to finding generalized eigenvectors $\vec{x}$ of matrices A and B satisfying $A\vec{x} = \lambda\vec{x}$ [70, 71]. Assume $A, B \in \mathbb{R}^{n \times n}$ are symmetric and positive definite.

(a) Define the *generalized Rayleigh quotient* $\rho(\vec{x})$ as the function

$$\rho(\vec{x}) \equiv \frac{\vec{x}^\top A\vec{x}}{\vec{x}^\top B\vec{x}}.$$

Show that $\nabla\rho$ is proportional to $A\vec{x} - \rho(\vec{x})B\vec{x}$.

(b) Show that critical points of $\rho(\vec{x})$ with $\vec{x} \neq \vec{0}$ are the generalized eigenvectors of (A, B). Argue that the largest and smallest generalized eigenvalues come from maximizing and minimizing $\rho(\vec{x})$, resp.

(c) If we search in the gradient direction from the current iterate $\vec{x}$, we must solve the following line search problem:

$$\min_{\alpha \in \mathbb{R}} \rho(\vec{x} + \alpha\vec{r}(\vec{x})),$$

where $r(\vec{x}) \equiv A\vec{x} - \rho(\vec{x})B\vec{x}$. Show that α can be found using the quadratic equation.

(d) Based on our construction above, propose an iteration for finding $\vec{x}$. When $B = I_{n \times n}$, is this method the same as the power method?

CHAPTER 11

Specialized Optimization Methods

CONTENTS

11.1 Nonlinear Least Squares 219
11.1.1 Gauss-Newton 220
11.1.2 Levenberg-Marquardt 220
11.2 Iteratively-Reweighted Least Squares 221
11.3 Coordinate Descent and Alternation 223
11.3.1 Identifying Candidates for Alternation 223
11.3.2 Augmented Lagrangians and ADMM 227
11.4 Global Optimization 232
11.4.1 Graduated Optimization 233
11.4.2 Randomized Global Optimization 235
11.5 Online Optimization 237

OUR initial methods for optimization comprised generic approaches to minimizing functions $f(\vec{x})$ with and without constraints on $\vec{x}$. These methods make few assumptions on the form of f or the constraints. When solving a linear system of equations, however, specialized iterative solvers like conjugate gradients have more reliable and efficient behavior than the generic tools.

In this chapter, we continue to exploit special problem structure to solve certain nonlinear problems more efficiently. Replacing monolithic generic algorithms with ones tailored to a given problem can make optimizations faster and easier to troubleshoot, although doing so can require more implementation effort than calling a pre-packaged solver.

11.1 NONLINEAR LEAST SQUARES

Recall the nonlinear regression problem posed in Example 8.1. If we wish to fit a function $y = ce^{ax}$ to a set of data points $(x_1, y_1), \ldots, (x_k, y_k)$, an optimization mimicking linear least-squares is to minimize the function

$$E(a, c) \equiv \sum_i (y_i - ce^{ax_i})^2.$$

This energy reflects the fact that we wish $y_i - ce^{ax_i} \approx 0$ for all i.

More generally, suppose we are given a set of functions $f_1(\vec{x}), \ldots, f_k(\vec{x})$ for $\vec{x} \in \mathbb{R}^n$. Then, if we want $f_i(\vec{x}) \approx 0$ for all i, then a reasonable objective trading off between these

terms is

$$E_{\text{NLS}}(\vec{x}) \equiv \frac{1}{2}\sum_i [f_i(\vec{x})]^2.$$

Objective functions of this form are known as *nonlinear least squares* problems. For the exponential regression problem above, we would take $f_i(a, c) \equiv y_i - ce^{ax_i}$.

11.1.1 Gauss-Newton

When we run Newton's method to minimize a function $f(\vec{x})$, we must know the gradient *and* Hessian of f. Knowing only the gradient of f is not enough, since approximating functions with planes provides no information about their extrema. Contrastingly, the Gauss-Newton algorithm for nonlinear least squares makes the observation that approximating each f_i with a linear function yields a nontrivial approximation of E_{NLS}. The main feature of this approach is that it requires only *first-order* approximation of the f_i's rather than Hessians.

Suppose we write

$$f_i(x) \approx f_i(\vec{x}_0) + [\nabla f_i(\vec{x}_0)] \cdot (\vec{x} - \vec{x}_0).$$

Then, we can approximate E_{NLS} with E^0_{NLS} given by

$$E^0_{\text{NLS}}(\vec{x}) = \frac{1}{2}\sum_i \left(f_i(\vec{x}_0) + [\nabla f_i(\vec{x}_0)] \cdot (\vec{x} - \vec{x}_0)\right)^2.$$

Define $F(\vec{x}) \equiv (f_1(\vec{x}), f_2(\vec{x}), \ldots, f_k(\vec{x}))$ by stacking the f_i's into a column vector. Then,

$$E^0_{\text{NLS}}(\vec{x}) = \frac{1}{2}\|F(\vec{x}_0) + DF(\vec{x}_0)(\vec{x} - \vec{x}_0)\|_2^2,$$

where DF is the Jacobian of F. Minimizing $E^0_{\text{NLS}}(\vec{x})$ is a *linear* least squares problem $-F(\vec{x}_0) \approx DF(\vec{x}_0)(\vec{x} - \vec{x}_0)$, which can be solved via the normal equations:

$$\vec{x} = \vec{x}_0 - (DF(\vec{x}_0)^\top DF(\vec{x}_0))^{-1} DF(\vec{x}_0)^\top F(\vec{x}_0).$$

More practically, the system can be solved using the QR factorization of $DF(\vec{x}_0)$ or—in higher dimensions—using conjugate gradients and related methods.

We can view $\vec{x}$ from minimizing $E^0_{\text{NLS}}(\vec{x})$ as an improved approximation of the minimum of $E_{\text{NLS}}(\vec{x})$ starting from $\vec{x}_0$. The Gauss-Newton algorithm iterates this formula to solve nonlinear least squares:

$$\vec{x}_{k+1} = \vec{x}_k - (DF(\vec{x}_k)^\top DF(\vec{x}_k))^{-1} DF(\vec{x}_k)^\top F(\vec{x}_k).$$

This iteration is not guaranteed to converge in all situations. Given an initial guess sufficiently close to the minimum of the nonlinear least squares problem, however, the approximation above behaves similarly to Newton's method and hence can have even quadratic convergence to the minimum.

11.1.2 Levenberg-Marquardt

The Gauss-Newton algorithm uses an approximation $E^0_{\text{NLS}}(\vec{x})$ of the nonlinear least-squares energy as a proxy for $E_{\text{NLS}}(\vec{x})$ that is easier to minimize. In practice, this approximation is likely to fail as $\vec{x}$ moves farther from $\vec{x}_0$, so we might modify the Gauss-Newton step to include a step size limitation:

$$\begin{array}{rl} \min_{\vec{x}} & E^0_{\text{NLS}}(\vec{x}) \\ \text{such that} & \|\vec{x} - \vec{x}_0\|_2^2 \leq \Delta \end{array}$$

That is, we now restrict our change in $\vec{x}$ to have norm less than some user-provided value Δ; the Δ neighborhood about $\vec{x}_0$ is called a *trust region*. Denote $H \equiv DF(\vec{x}_0)^\top DF(\vec{x}_0)$ and $\delta\vec{x} \equiv \vec{x} - \vec{x}_0$. Then, we can solve:

$$\begin{array}{rl} \min_{\vec{x}} & \frac{1}{2}\delta\vec{x}^\top H\delta\vec{x} + F(\vec{x}_0)^\top DF(\vec{x}_0)\delta\vec{x} \\ \text{such that} & \|\delta\vec{x}\|_2^2 \leq \Delta \end{array}$$

That is, we displace $\vec{x}$ by minimizing the Gauss-Newton approximation after imposing the step size restriction. This problem has the following KKT conditions (see §9.2.2):

$$\begin{aligned} \text{Stationarity: } \vec{0} &= H\delta\vec{x} + DF(\vec{x}_0)^\top F(\vec{x}_0) + 2\mu\delta\vec{x} \\ \text{Primal feasibility: } \|\delta\vec{x}\|_2^2 &\leq \Delta \\ \text{Complementary slackness: } \mu(\Delta - \|\delta x\|_2^2) &= 0 \\ \text{Dual feasibility: } \mu &\geq 0 \end{aligned}$$

Define $\lambda \equiv 2\mu$. Then, the stationarity condition can be written as follows:

$$(H + \lambda I_{n\times n})\delta\vec{x} = -DF(\vec{x}_0)^\top F(\vec{x}_0)$$

Assume the constraint $\|\delta\vec{x}\|_2 \leq \Delta$ is active. Then, since $\lambda > 0$ and H is positive semidefinite, we know $H + \lambda I_{n\times n}$ must be positive definite.

Based on this formula, the Levenberg-Marquardt algorithm makes the following step [77, 80]:

$$\vec{x} = \vec{x}_0 - (DF(\vec{x}_0)^\top DF(\vec{x}_0) + \lambda I_{n\times n})^{-1} F(\vec{x}_0)$$

This linear system can also be derived by applying Tikhonov regularization to the Gauss-Newton linear system. When λ is small, it becomes similar to the Gauss-Newton algorithm, while large λ corresponds to a gradient descent step on E_{NLS}.

Rather than specifying Δ as introduced above, Levenberg-Marquardt steps work with $\lambda > 0$ directly. By the KKT conditions, *a posteriori* we know this choice corresponds to having taken $\Delta = \|\vec{x} - \vec{x}_0\|_2^2$. As $\lambda \to \infty$, the step from Levenberg-Marquardt satisfies $\|\vec{x} - \vec{x}_0\| \to 0$; so, we can regard Δ and λ as approximately inversely proportional.

Typical approaches adaptively adjust the *damping parameter* λ during each each iteration:

$$\vec{x}_{k+1} = \vec{x}_k - (DF(\vec{x}_k)^\top DF(\vec{x}_k) + \lambda_k I_{n\times n})^{-1} F(\vec{x}_k)$$

For instance, we can scale up λ_k when the step in $F(\vec{x})$ agrees well with the approximate value predicted by $E^0_{\text{NLS}}(\vec{x})$, since this corresponds to increasing the size of the neighborhood in which the Gauss-Newton approximation is effective. The quality of the approximation also can be improved by replacing $I_{n\times n}$ with the diagonal of H, which helps account for the curvature of the objective function. Many of these methods are heuristic, but they dramatically improve convergence rates.

11.2 ITERATIVELY-REWEIGHTED LEAST SQUARES

Continuing in our consideration of least-squares problems, suppose we wish to minimize a function of the form:

$$E_{\text{IRLS}}(\vec{x}) \equiv \sum_i f_i(\vec{x})[g_i(\vec{x})]^2$$

We can think of $f_i(\vec{x})$ as a *weight* on the least-squares term $g_i(\vec{x})$.

Example 11.1 (L^p optimization)**.** Similar to the compressed sensing problems in §9.4.1, given $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$ we can generalize least-squares by minimizing

$$E_p(\vec{x}) \equiv \|A\vec{x} - \vec{b}\|_p^p.$$

Choosing $p = 1$ can promote sparsity in the residual $\vec{b} - A\vec{x}$. We can write this function in an alternative form:

$$E_p(\vec{x}) = \sum_i (\vec{a}_i \cdot \vec{x} - b_i)^{p-2}(\vec{a}_i \cdot \vec{x} - b_i)^2.$$

Here, we denote the rows of A as $\vec{a}_i^\top$. Then, $E_p = E_{\text{IRLS}}$ after defining:

$$f_i(\vec{x}) = (\vec{a}_i \cdot \vec{x} - b_i)^{p-2}$$
$$g_i(\vec{x}) = \vec{a}_i \cdot \vec{x} - b_i$$

The *iteratively-reweighted least squares* (IRLS) algorithm makes use of the following fixed-point iteration:

$$\vec{x}_{k+1} = \min_{\vec{x}_{k+1}} \sum_i f_i(\vec{x}_k)[g_i(\vec{x}_{k+1})]^2$$

In the minimization, $\vec{x}_k$ is fixed, so the optimization is a least-squares problem over the g_i's. When g_i is linear, the minimization can be carried out via a matrix inverse; otherwise we can use the nonlinear least-squares techniques in §11.1.

Example 11.2 (L^1 optimization)**.** Continuing Example 11.1, suppose we take $p = 1$. Then,

$$E_1(\vec{x}) = \sum_i |\vec{a}_i \cdot \vec{x} - b_i| = \sum_i \frac{1}{|\vec{a}_i \cdot \vec{x} - b_i|}(\vec{a}_i \cdot \vec{x} - b_i)^2.$$

This functional leads to the following IRLS iteration:

$w_i \leftarrow [\max(\lvert\vec{a}_i \cdot \vec{x} - b_i\rvert, \delta)]^{-1}$	▷ Recompute weights
$\vec{x} \leftarrow \min_{\vec{x}} \sum_i w_i(\vec{a}_i \cdot \vec{x} - b_i)^2$	▷ Linear least-squares

The parameter $\delta > 0$ avoids division by zero; large values of δ make better-conditioned linear systems but worse approximations of the original $\|\cdot\|_1$ problem.

Example 11.3 (Weiszfeld algorithm)**.** Recall the geometric median problem from Example 8.3. In this problem, given $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$ we wish to minimize

$$E(\vec{x}) \equiv \sum_i \|\vec{x} - \vec{x}_i\|_2.$$

Similar to the L^1 problem in Example 11.2, we can write this function as a weighted least-squares problem:

$$E(\vec{x}) \equiv \sum_i \frac{1}{\|\vec{x} - \vec{x}_i\|_2}\|\vec{x} - \vec{x}_i\|_2^2.$$

Then, IRLS provides the *Weiszfeld algorithm* for geometric median problems:

$w_i \leftarrow [\max(\|\vec{x} - \vec{x}_i\|_2, \delta)]^{-1}$	▷ Recompute weights
$\vec{x} \leftarrow \min_{\vec{x}} \sum_i w_i(\vec{x} - \vec{x}_i)^2$	▷ Linear least-squares

We can solve for the second step of the Weiszfeld algorithm in closed form. Differentiating the objective with respect to $\vec{x}$ shows

$$\vec{0} = \sum_i 2w_i(\vec{x} - \vec{x}_i) \implies \vec{x} = \frac{\sum_i w_i \vec{x}_i}{\sum_i w_i}.$$

Thus, the two alternating steps of Weiszfeld's algorithm can be carried out efficiently as:

$$\begin{array}{ll} w_i \leftarrow [\max(\|\vec{x} - \vec{x}_i\|_2, \delta)]^{-1} & \triangleright \text{ Recompute weights} \\ \vec{x} \leftarrow \frac{\sum_i w_i \vec{x}_i}{\sum_i w_i} & \triangleright \text{ Weighted centroid} \end{array}$$

IRLS algorithms are straightforward to formulate, so they are worth trying if an optimization can be written in the form of E_{IRLS}. In particular, when g_i is linear for all i as in Example 11.2, each iteration of IRLS can be carried out quickly using Cholesky factorization, QR, conjugate gradients, and so on, avoiding line search and other more generic strategies.

It is difficult to formulate general conditions under which IRLS will reach the minimum of E_{IRLS}. Often iterates must be approximated somewhat as in the introduction of δ to Example 11.2 to avoid division by zero and other degeneracies. In the case of L^1 optimization, however, IRLS can be shown with small modification to converge to the optimal point [29].

11.3 COORDINATE DESCENT AND ALTERNATION

Suppose we wish to minimize a function $f : \mathbb{R}^{n+m} \to \mathbb{R}$. Rather than viewing the input as a single variable $\vec{x} \in \mathbb{R}^{n+m}$, we might write f in an alternative form as $f(\vec{x}, \vec{y})$, for $\vec{x} \in \mathbb{R}^n$ and $\vec{y} \in \mathbb{R}^m$. One strategy for optimization is to fix $\vec{y}$ and minimize f with respect to $\vec{x}$, fix $\vec{x}$ and minimize f with respect to $\vec{y}$, and repeat:

$$\begin{array}{ll} \textbf{for } i \leftarrow 1, 2, \ldots & \\ \quad \vec{x}_{i+1} \leftarrow \min_{\vec{x}} f(\vec{x}, \vec{y}_i) & \triangleright \text{ Optimize } \vec{x} \text{ with } \vec{y} \text{ fixed} \\ \quad \vec{y}_{i+1} \leftarrow \min_{\vec{y}} f(\vec{x}_{i+1}, \vec{y}) & \triangleright \text{ Optimize } \vec{y} \text{ with } \vec{x} \text{ fixed} \end{array}$$

In this *alternating* approach, the value of $f(\vec{x}_i, \vec{y}_i)$ decreases monotonically as i increases since a minimization is carried out at each step. We cannot prove that alternation always reaches a global or even local minimum, but in many cases it can be an efficient option for otherwise challenging problems.

11.3.1 Identifying Candidates for Alternation

There are a few reasons why we might wish to perform alternating optimization:

- The individual problems over $\vec{x}$ and $\vec{y}$ are optimizations in a lower dimension and may converge more quickly.
- We may be able to split the variables in such a way that the individual $\vec{x}$ and $\vec{y}$ steps are far more efficient than optimizing both variables jointly.

Below we provide a few examples of alternating optimization in practice.

Example 11.4 (Generalized PCA)**.** In the PCA problem from §6.2.5, we are given a data matrix $X \in \mathbb{R}^{n\times k}$ whose columns are k data points in $\mathbb{R}^n$. We seek a basis in $\mathbb{R}^n$ of size d such that the projection of the data points onto the basis introduces minimal approximation error; we will store this basis in the columns of $C \in \mathbb{R}^{n\times d}$. Classical PCA minimizes $\|X - CY\|_{\text{Fro}}^2$ over both C and Y, where the columns of $Y \in \mathbb{R}^{d\times k}$ are the coefficients of the approximations of our data points in the C basis. If C is constrained to be orthogonal, then $Y = C^\top X$, recovering the formula in our previous discussion.

The Frobenius norm in PCA is somewhat arbitrary: The relevant relationship is $X - CY \approx 0$. Alternative PCA models minimize $\mu(X - CY)$ over C and Y, for some other energy function $\mu : \mathbb{R}^{n\times k} \to \mathbb{R}$ favoring matrices with entries near zero; μ can provide enhanced robustness to noise or encode application-specific assumptions. Taking $\mu(M) \equiv \|M\|_{\text{Fro}}^2$ recovers classical PCA; another popular choice is *robust PCA*, which takes $\mu(M) \equiv \sum_{ij} |M_{ij}|$ [66].

The product CY in $\mu(X-CY)$ makes the energy nonlinear and nonconvex. A typical minimization routine for this problem uses alternation: First optimize C with Y fixed, then optimize Y with C fixed, and repeat. Whereas optimizing the energy with respect to C and Y jointly might require a generic large-scale method, the individual alternating C and Y steps can be easier:

- When $\mu(M) = \|M\|_{\text{Fro}}^2$, the Y and C alternations each are least-squares problems, leading to the *alternating least-squares* (ALS) algorithm for classical PCA.
- When $\mu(M) \equiv \sum_{ij} |M_{ij}|$, the Y and C alternations are linear programs, which can be optimized using the techniques mentioned in §9.4.1.

Example 11.5 (ARAP)**.** Recall the planar "as-rigid-as-possible" (ARAP) energy introduced in Example 9.5:

$$\begin{aligned}\text{minimize}_{R_v,\vec{y}_v} &\sum_{v\in V}\sum_{(v,w)\in E} \|R_v(\vec{x}_v - \vec{x}_w) - (\vec{y}_v - \vec{y}_w)\|_2^2\\ \text{such that } &R_v^\top R_v = I_{2\times 2}\ \forall v \in V\\ &\vec{y}_v \text{ fixed } \forall v \in V_0\end{aligned}$$

Solving for the matrices $R_v \in \mathbb{R}^{2\times 2}$ and vertex positions $\vec{y}_v \in \mathbb{R}^2$ simultaneously is a highly nonlinear and nonconvex task, especially given the orthogonality constraint $R_v^\top R_v = I_{2\times 2}$. There is one $\vec{y}_v$ and one R_v for each vertex v of a triangle mesh with potentially thousands or even millions of vertices, so such a direct optimization using quasi-Newton methods requires a large-scale linear solve per iteration and still is prone to finding only local minima.

Instead, [110] suggests an alternating approach:

- Fixing the R_v matrices and optimizing only for the positions $\vec{y}_v$ is a least-squares problem:

$$\begin{aligned}\text{minimize}_{\vec{y}_v} &\sum_{v\in V}\sum_{(v,w)\in E} \|R_v(\vec{x}_v - \vec{x}_w) - (\vec{y}_v - \vec{y}_w)\|_2^2\\ \text{such that } &\vec{y}_v \text{ fixed } \forall v \in V_0\end{aligned}$$

  Optimizing for the $\vec{y}_v$'s with the remaining variables fixed is a sparse, positive-definite linear solve.

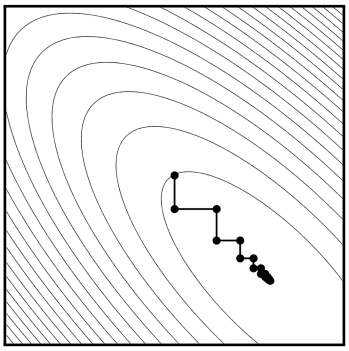


FIGURE 11.1 Coordinate descent in two dimensions alternates between minimizing in the horizontal and vertical axis directions.

- Now, suppose we fix the $\vec{y}_v$'s and optimize for the R_v's. In this step, no energy terms or constraints couple any pair R_v, R_w for $v, w \in V$, so we can solve for each matrix R_v independently. That is, rather than solving for $4|V|$ unknowns simultaneously, we can loop over $v \in V$ and in each step solve the following optimization for a single $\mathbb{R}_v \in \mathbb{R}^{2\times 2}$:

$$\begin{aligned} &\text{minimize}_{R_v} \sum_{(v,w)\in E} \|R_v(\vec{x}_v - \vec{x}_w) - (\vec{y}_v - \vec{y}_w)\|_2^2 \\ &\text{such that } R_v^\top R_v = I_{2\times 2} \;\forall v \in V \end{aligned}$$

  This optimization is an instance of the Procrustes problem from §6.2.4 and hence can be carried out in closed-form using a 2×2 SVD. We have replaced a large-scale minimization step with the application of a formula that can be evaluated in parallel for each vertex, a massive computational savings.

Alternating between optimizing for the $\vec{y}_v$'s with the R_v's fixed and vice versa decreases the energy using two efficient tools, a linear solve and a collection of 2×2 SVD factorizations. This is far more efficient than considering all the variables simultaneously, and in practice a few iterations of this alternation can be sufficient to generate elastic deformations like the one shown in Figure 9.3. Extensions of ARAP even run in real time, optimizing the energy fast enough to provide interactive feedback to artists editing two- and three-dimensional shapes.

Example 11.6 (Coordinate descent)**.** Taking the philosophy of alternating optimization to an extreme, rather than splitting the inputs of $f : \mathbb{R}^n \to \mathbb{R}$ into two variables, we could view f as a function of several variables $f(x_1, x_2, \ldots, x_n)$. Then, we could cycle through each input x_i and perform a one-dimensional optimization. This lightweight algorithm, illustrated in Figure 11.1, is known as *coordinate descent*.

For instance, suppose we wish to solve the least-squares problem $A\vec{x} \approx \vec{b}$ by minimizing $\|A\vec{x} - \vec{b}\|_2^2$. Just like in Chapter 10, the line search over any single x_i can be solved in closed form. In particular, if the columns of A are vectors $\vec{a}_1, \ldots, \vec{a}_n$, then as shown in §0.3.1 we can write $A\vec{x} - \vec{b} = x_1\vec{a}_1 + \cdots + x_n\vec{a}_n - \vec{b}$. By this expansion, we know:

$$0 = \frac{\partial}{\partial x_i}\|x_1\vec{a}_1 + \cdots + x_n\vec{a}_n - \vec{b}\|_2^2 = 2(A\vec{x} - \vec{b}) \cdot \vec{a}_i = \sum_j \left[\left(\sum_k a_{ji}a_{jk}x_k\right) - a_{ji}b_j\right]$$

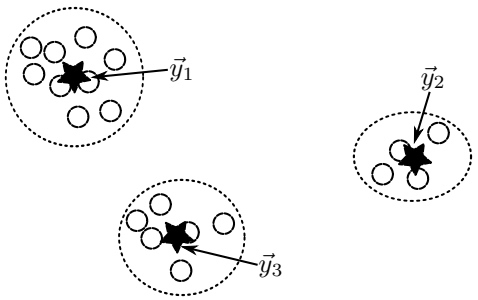


FIGURE 11.2 The k-means algorithm seeks cluster centers $\vec{y}_i$ that partition a set of data points $\vec{x}_1, \ldots, \vec{x}_m$ based on their closest center.

Solving this expression for x_i yields the following coordinate descent update for a single x_i:

$$x_i \leftarrow \frac{\vec{a}_i \cdot \vec{b} - \sum_{k \neq i} x_k (\vec{a}_i \cdot \vec{a}_k)}{\|\vec{a}_i\|_2^2}$$

Coordinate descent for least-squares iterates this formula over $i = 1, 2, \ldots, n$ repeatedly until convergence. This approach has efficient localized updates and appears in machine learning methods where A has many more rows than columns, sampled from a data distribution. We have traded a global method for one that locally updates the solution $\vec{x}$ by solving extremely simple subproblems.

Example 11.7 (k-means clustering)**.** Suppose we are given a set of data points $\vec{x}_1, \ldots, \vec{x}_m \in \mathbb{R}^n$ and wish to group these points into k clusters based on distance, as illustrated in Figure 11.2. Take $\vec{y}_1, \ldots, \vec{y}_k \in \mathbb{R}^n$ to be the centers of clusters $1, \ldots, k$, respectively. Then, the k-means technique attempts to minimize the following energy:

$$E(\vec{y}_1, \ldots, \vec{y}_k) \equiv \sum_{i=1}^{m} \min_{c \in \{1,\ldots,k\}} \|\vec{x}_i - \vec{y}_c\|_2^2$$

In words, E measures the total squared distance of the data points $\vec{x}_i$ to their closest cluster center $\vec{y}_c$.

Define $c_i \equiv \arg\min_{c \in \{1,\ldots,k\}} \|\vec{x}_i - \vec{y}_c\|_2^2$; that is, c_i is the index of the cluster center $\vec{y}_{c_i}$ closest to $\vec{x}_i$. From this substitution, we can write an equivalent formulation of the k-means objective as follows:

$$E(\vec{y}_1, \ldots, \vec{y}_k; c_1, \ldots, c_m) \equiv \sum_{i=1}^{m} \|\vec{x}_i - \vec{y}_{c_i}\|_2^2$$

The variables c_i are integers, but we can still optimize using alternation:

- When the c_i's are fixed, the optimization for the $\vec{y}_j$'s is a least-squares solve that can be written in closed form:

$$\vec{y}_j = \frac{\sum_{c_i = j} \vec{x}_i}{|\{c_i = j\}|}$$

That is, $\vec{y}_j$ is the average of the points $\vec{x}_i$ assigned to cluster j.

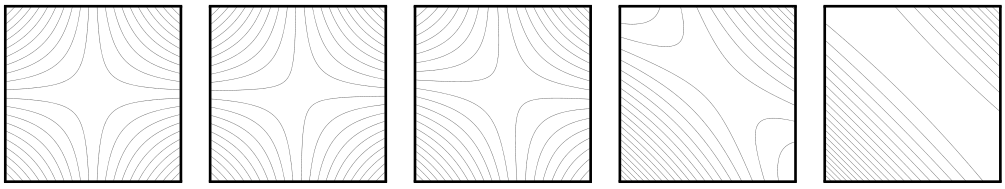


FIGURE 11.3 We can minimize $f(x, y) \equiv xy$ subject to $x + y = 1$ approximately by minimizing the quadratically-penalized version $f_\rho(x, y) = xy + \rho(x + y - 1)^2$. As ρ increases, however, the level sets of xy get obscured in favor of enforcing the constraint.

- The optimization for c_i can be carried out using the expression $c_i \equiv \arg\min_{c \in \{1,\ldots,k\}} \|\vec{x}_i - \vec{y}_c\|_2^2$ by iterating from 1 to k for each i.

This alternation is known as the k-means algorithm and is one of the most well-known methods for clustering. One drawback of this method is that it is highly sensitive to the initial guesses of $\vec{y}_1, \ldots, \vec{y}_k$. In practice, k-means is run several times with different initial guesses and only the best output is preserved. Alternatively, methods like "k-means++" specifically design initial values of the $\vec{y}_i$'s to encourage convergence to a better local minimum [3].

11.3.2 Augmented Lagrangians and ADMM

Nonlinear constrained problems are often the most challenging optimization tasks. While the general algorithms in §9.3 are applicable, they can be sensitive to the initial guess of the minimizer, slow to iterate due to large linear solves, and slow to converge in the absence of more information about the problems at hand. Using these methods is easy from an engineering perspective since they require providing only a function and its derivatives, but with some additional work on paper, certain objective functions can be tackled using faster techniques, many of which can be parallelized on multiprocessor machines. It is worth checking if a problem can be solved via one of these strategies, especially when the dimensionality is high or the objective has a number of similar or repeated terms.

In this section, we consider an alternating approach to equality-constrained optimization that has gained considerable attention in recent literature. While it can be used out-of-the-box as yet another generic optimization algorithm, its primary value appears to be in the decomposition of complex minimization problems into simpler steps that can be iterated, often in parallel. In large part we will follow the development of [12], which contains many examples of applications of this class of techniques.

As stated in Chapter 9, the equality-constrained optimization problem can be stated as follows:

$$\begin{aligned} &\text{minimize } f(\vec{x}) \\ &\text{such that } g(\vec{x}) = \vec{0} \end{aligned}$$

The approach suggested in §9.3.2 optimizes an unconstrained objective with a quadratic penalty:

$$f_\rho(\vec{x}) = f(\vec{x}) + \frac{1}{2}\rho\|g(\vec{x})\|_2^2$$

As $\rho \to \infty$, critical points of f_ρ satisfy the $g(\vec{x}) = \vec{0}$ constraint more and more strongly. The trade-off for this method, however, is that the optimization becomes more and more poorly-conditioned as ρ becomes large. This effect is illustrated in Figure 11.3; when ρ is large, the level sets of f_ρ mostly are dedicated to enforcing the constraint rather than minimizing the original function $f(\vec{x})$, making it difficult to distinguish between potential $\vec{x}$'s that all satisfy the constraint.

Alternatively, by the method of Lagrange multipliers (Theorem 0.1), we can seek first-order optima of this problem as the critical points of $\Lambda(\vec{x}, \vec{\lambda})$ given by

$$\Lambda(\vec{x}, \vec{\lambda}) \equiv f(\vec{x}) - \vec{\lambda}^\top g(\vec{x})$$

This Lagrangian does not suffer from conditioning issues that affect the quadratic penalty method. On the other hand, it replaces a minimization problem—which can be solved by moving "downhill"—with a more challenging saddle point problem in which critical points should be minima of Λ with respect to $\vec{x}$ and maxima of Λ with respect to $\vec{\lambda}$. Optimizing by alternatively minimizing with respect to $\vec{x}$ and maximizing with respect to $\vec{\lambda}$ can be unstable; intuitively this makes some sense since the two iterations make it unclear whether Λ should be small or large.

The *augmented Lagrangian* method for equality-constrained optimization combines the quadratic penalty and Lagrangian strategies, using the penalty to "soften" individual iterations of the alternation for optimizing Λ described above. It replaces the original equality-constrained optimization problem with the following equivalent augmented problem:

$$\begin{aligned} &\text{minimize } f(\vec{x}) + \frac{1}{2}\rho\|g(\vec{x})\|_2^2 \\ &\text{such that } g(\vec{x}) = \vec{0} \end{aligned}$$

Any $\vec{x}$ satisfying the $g(\vec{x}) = \vec{0}$ constraint makes the second objective term vanish. But, when the constraint is not exactly satisfied, the second energy term biases the objective toward points $\vec{x}$ that approximately satisfy the equality constraint. In other words, during iterations of augmented Lagrangian optimization, the $\rho\|g(\vec{x})\|_2^2$ acts as a rubber band pulling $\vec{x}$ closer to the constraint set even during the minimization step.

This modified problem has a new Lagrangian given by:

$$\Lambda_\rho(\vec{x}, \vec{\lambda}) \equiv f(\vec{x}) + \frac{1}{2}\rho\|g(\vec{x})\|_2^2 - \vec{\lambda}^\top g(\vec{x})$$

Hence, the augmented Lagrangian method optimizes this objective by alternating as follows:

```
for i ← 1, 2, ...
    λ_{i+1} ← λ_i − ρ g(x_i)              ▷ Dual update
    x_i ← min_x Λ_ρ(x, λ_{i+1})           ▷ Primal update
```

The dual update step can be thought of as a gradient ascent step for $\vec{\lambda}$. The parameter ρ here no longer has to approach infinity for convergence, since the Lagrange multiplier enforces the constraint regardless; instead, the quadratic penalty serves to make sure the output of the $\vec{x}$ iteration does not violate the constraints too strongly.

Augmented Lagrangian optimization has the advantage that it alternates between applying a formula to update $\vec{\lambda}$ and solving an *unconstrained* minimization problem for $\vec{x}$. For many optimization problems, however, the unconstrained objective still may be non-differentiable or difficult to optimize. A few special cases, e.g. Uzawa iteration for dual

decomposition [118], can provide methods for optimization but in many circumstances quasi-Newton methods outperform this approach with limited loss of efficiency and more favorable convergence guarantees.

A small alteration to general augmented Lagranian minimization, however, yields the *alternating direction method of multipliers* (ADMM) for optimizing slightly more specific problems of the form:

$$\begin{aligned} &\text{minimize } f(\vec{x}) + h(\vec{z}) \\ &\text{such that } A\vec{x} + B\vec{z} = \vec{c} \end{aligned}$$

Here, we optimize over both $\vec{x}$ and $\vec{z}$, where $f, h : \mathbb{R}^n \to \mathbb{R}$ are given functions and the equality constraint is linear. As we will show, this form encodes many common classes of optimization problems. We will design an algorithm that carries out alternation between the two primal variables $\vec{x}$ and $\vec{z}$, as well as between primal and dual optimization.

The augmented Lagrangian in this case is:

$$\Lambda_\rho(\vec{x}, \vec{z}, \vec{\lambda}) \equiv f(\vec{x}) + h(\vec{z}) + \frac{1}{2}\rho\|A\vec{x} + B\vec{z} - \vec{c}\|_2^2 + \vec{\lambda}^\top(A\vec{x} + B\vec{z} - \vec{c})$$

Alternating between optimizing $\vec{x}$, $\vec{z}$, and $\vec{\lambda}$ leads to a modification of the augmented Lagrangian method:

for $i \leftarrow 1, 2, \ldots$
- $\vec{x}_{i+1} \leftarrow \arg\min_{\vec{x}} \Lambda_\rho(\vec{x}, \vec{z}_i, \vec{\lambda}_i)$ ▷ $\vec{x}$ update
- $\vec{z}_{i+1} \leftarrow \arg\min_{\vec{z}} \Lambda_\rho(\vec{x}_{i+1}, \vec{z}, \vec{\lambda}_i)$ ▷ $\vec{z}$ update
- $\vec{\lambda}_{i+1} \leftarrow \vec{\lambda}_i + \rho(A\vec{x}_{i+1} + B\vec{z}_{i+1} - \vec{c})$ ▷ Dual update

In this algorithm, $\vec{x}$ and $\vec{z}$ are optimized one-at-a-time; the augmented Lagrangian method optimizes them jointly. Although this splitting can require more iterations for convergence, clever choices of $\vec{x}$ and $\vec{z}$ lead to powerful division-of-labor strategies for breaking down difficult minimizations. Each individual iteration will take *far* less time, even though more iterations may be needed for convergence.

In a sense, ADMM is a "meta-algorithm" used to design optimization techniques. Rather than calling a generic package to minimize Λ_ρ with respect to $\vec{x}$ and $\vec{z}$, we will find choices of $\vec{x}$ and $\vec{z}$ that make individual steps fast.

Before working out several examples of ADMM in action, it is worth noting that it is guaranteed to converge to a critical point of the objective under fairly weak conditions. For instance, ADMM reaches a global minimum when f and h are convex and Λ_ρ has a saddle point. ADMM has also been observed to converge even for nonconvex problems, although theoretical understanding in this case is limited. In practice, ADMM tends to be quick to generate *approximate* minima of the objective but can require a long tail of iterations to squeeze out the last decimal points of accuracy; for this reason, some systems use ADMM to do large-scale steps and transition to other strategies for localized optimization.

We dedicate the remainder of this section to working out some examples of ADMM in practice. The general pattern here is to split the optimization variables into $\vec{x}$ and $\vec{z}$ in such a way that the two primal update steps each can be carried out efficiently, preferably in closed form or decoupling in such a way that parallelized computations can be used to solve many subproblems at once. This makes individual iterations of ADMM inexpensive.

Example 11.8 (Nonnegative least-squares)**.** Suppose we wish to minimize $\|A\vec{x} - \vec{b}\|_2^2$ with respect to $\vec{x}$ subject to the constraint $\vec{x} \geq \vec{0}$. The $\vec{x} \geq 0$ constraint rules out using Gaussian elimination on the normal equations, but ADMM provides one method to bypass this issue.

Consider solving the following equivalent problem:

$$\begin{aligned} &\text{minimize } \|A\vec{x} - \vec{b}\|_2^2 + h(\vec{z}) \\ &\text{such that } \vec{x} = \vec{z} \end{aligned}$$

Here, we define the new function $h(\vec{z})$ as follows:

$$h(\vec{z}) = \begin{cases} 0 & \vec{z} \geq \vec{0} \\ \infty & \text{otherwise} \end{cases}$$

The function $h(\vec{z})$ is discontinuous, but it is convex. The minimization over $\vec{x}$ and $\vec{z}$ jointly solves the original nonnegative least-squares problem after applying the $\vec{x} = \vec{z}$ constraint.

The augmented Lagrangian in this case is:

$$\Lambda_\rho(\vec{x}, \vec{z}, \vec{\lambda}) = \|A\vec{x} - \vec{b}\|_2^2 + h(\vec{z}) + \frac{1}{2}\rho\|\vec{x} - \vec{z}\|_2^2 + \vec{\lambda}^\top(\vec{x} - \vec{z})$$

If we fix $\vec{z}$, the Λ_ρ is differentiable. Hence, we can carry out the $\vec{x}$ iteration by writing the following:

$$\begin{aligned} \vec{0} &= \nabla_{\vec{x}}\Lambda_\rho(\vec{x}, \vec{z}, \vec{\lambda}) \\ &= 2A^\top A\vec{x} - 2A^\top\vec{b} + \rho(\vec{x} - \vec{z}) + \vec{\lambda} \\ &= (2A^\top A + \rho I_{n\times n})\vec{x} + (\vec{\lambda} - 2A^\top\vec{b} - \rho\vec{z}) \\ \implies \vec{x} &= (2A^\top A + \rho I_{n\times n})^{-1}(2A^\top\vec{b} + \rho\vec{z} - \vec{\lambda}) \end{aligned}$$

This linear solve is a Tikhonov-regularized least-squares problem.

Minimizing Λ_ρ with respect to $\vec{z}$ also can be carried out in closed form. In particular, any objective function involving h effectively constrains each component of $\vec{z}$ to be nonnegative, so we can solve the $\vec{z}$ step using the following optimization:

$$\begin{aligned} &\text{minimize}_{\vec{z}} \ \frac{1}{2}\rho\|\vec{x} - \vec{z}\|_2^2 + \vec{\lambda}^\top(\vec{x} - \vec{z}) \\ &\text{such that } \vec{z} \geq \vec{0} \end{aligned}$$

The $\|A\vec{x} - \vec{b}\|_2^2$ is removed because it has no $\vec{z}$ dependence. This problem *decouples* over the components of $\vec{z}$ since no energy terms involve more than one dimension of $\vec{z}$ at a time. So, we can solve many instances of the following one-dimensional problem:

$$\begin{aligned} &\text{minimize}_{z_i} \ \frac{1}{2}\rho(x_i - z_i)^2 + \lambda_i(x_i - z_i) \\ &\text{such that } z_i \geq 0 \end{aligned}$$

In the absence of the $z_i \geq 0$ constraint, the objective is minimized when $0 = \rho(z_i - x_i) - \lambda_i \implies z_i = {}^{\lambda_i}/_{\rho} + x_i$; when this value is negative we instead take $z_i = 0$.

Hence, the ADMM algorithm for nonnegative least-squares is:

for $i \leftarrow 1, 2, \ldots$
  $\vec{x}_{i+1} \leftarrow (2A^\top A + \rho I_{n\times n})^{-1}(2A^\top \vec{b} + \rho \vec{z}_i - \vec{\lambda}_i)$ ▷ $\vec{x}$ update; least-squares
  $\vec{z}^0 \leftarrow \vec{\lambda}_i/\rho + \vec{x}_{i+1}$ ▷ Unconstrained $\vec{z}$ formula
  $\vec{z}_{i+1} \leftarrow$ Elementwise-Max$(\vec{z}^0, \vec{0})$ ▷ Enforce $\vec{z} \geq \vec{0}$
  $\vec{\lambda}_{i+1} \leftarrow \vec{\lambda}_i + \rho(\vec{x}_{i+1} - \vec{z}_{i+1})$ ▷ Dual update

This ADMM algorithm for nonnegative least-squares took our original problem—a quadratic program that could require difficult constrained optimization techniques–and replaced it with an alternation between a linear solve for $\vec{x}$, a formula for $\vec{z}$, and a formula for $\vec{\lambda}$. These individual steps are straightforward to implement and efficient to carry out.

Example 11.9 (ADMM for geometric median)**.** Returning to Example 11.3, we can reconsider the energy $E(\vec{x})$ for the geometric median problem using the machinery of ADMM:

$$E(\vec{x}) \equiv \sum_{i=1}^{N} \|\vec{x} - \vec{x}_i\|_2.$$

This time, we will split the problem into two unknowns $\vec{z}_i, \vec{x}$:

$$\text{minimize } \sum_i \|\vec{z}_i\|_2$$
$$\text{such that } \vec{z}_i + \vec{x} = \vec{x}_i \ \forall i$$

The augmented Lagrangian for this problem is:

$$\Lambda_\rho = \sum_i \left[\|\vec{z}_i\|_2 + \frac{1}{2}\rho\|\vec{z}_i + \vec{x} - \vec{x}_i\|_2^2 + \vec{\lambda}_i^\top(\vec{z}_i + \vec{x} - \vec{x}_i) \right]$$

As a function of $\vec{x}$, the augmented Lagrangian is differentiable and hence to find the $\vec{x}$ iteration we write:

$$\vec{0} = \nabla_{\vec{x}}\Lambda_\rho = \sum_i \left[\rho(\vec{x} - \vec{x}_i + \vec{z}_i) + \vec{\lambda}_i \right]$$
$$\implies \vec{x} = \frac{1}{N}\sum_i \left[\vec{x}_i - \vec{z}_i - \frac{1}{\rho}\vec{\lambda}_i \right]$$

The optimization for $\vec{z}_i$ decouples over i, so after removing terms that are constants in this step we wish to minimize $\|\vec{z}_i\|_2 + \frac{1}{2}\rho\|\vec{z}_i + \vec{x} - \vec{x}_i\|_2^2 + \vec{\lambda}_i^\top \vec{z}_i$ over $\vec{z}_i$ for each i separately. We can simplify the quadratic term as follows:

$$\frac{1}{2}\rho\|\vec{z}_i + \vec{x} - \vec{x}_i\|_2^2 + \vec{\lambda}_i^\top \vec{z}_i = \frac{1}{2}\rho\|\vec{z}_i\|_2^2 + \rho\vec{z}_i^\top\left(\frac{1}{\rho}\vec{\lambda}_i + \vec{x} - \vec{x}_i\right) + \text{const.}$$
$$= \frac{1}{2}\rho\left\|\vec{z}_i + \frac{1}{\rho}\vec{\lambda}_i + \vec{x} - \vec{x}_i\right\|_2^2 + \text{const.}$$

The constant terms can have $\vec{x}$ dependence since it is fixed in the $\vec{z}_i$ iteration. Defining $\vec{z}^0 \equiv -\frac{1}{\rho}\vec{\lambda}_i - \vec{x} + \vec{x}_i$, in the $\vec{z}_i$ iteration we have shown that we can solve:

$$\min_{\vec{z}_i}\left[\|\vec{z}_i\|_2 + \frac{1}{2}\rho\|\vec{z}_i - \vec{z}^0\|_2^2 \right].$$

Written in this form, it is clear that the optimal $\vec{z}_i$ satisfies $\vec{z}_i = t\vec{z}^0$ for some $t \in [0, 1]$, since the two terms of the objective trade off distances of $\vec{z}_i$ to $\vec{0}$ and $\vec{z}^0$. Hence, after dividing by $\|\vec{z}^0\|_2$ we can solve:

$$\min_{t \geq 0} \left[t + \frac{1}{2}\rho\|\vec{z}^0\|_2(t-1)^2 \right]$$

Using elementary calculus techniques we find:

$$t = \begin{cases} 1 - 1/\rho\|\vec{z}^0\|_2 & \text{when } \rho\|\vec{z}^0\|_2 \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Taking $\vec{z}_i = t\vec{z}_0$ finishes the $\vec{z}$ iteration of ADMM.

In summary, the ADMM algorithm for geometric median computation is as follows:

for $i \leftarrow 1, 2, \ldots$
  $\vec{x} \leftarrow \frac{1}{N}\sum_i \left[\vec{x}_i - \vec{z}_i - \frac{1}{\rho}\vec{\lambda}_i \right]$ ▷ $\vec{x}$ update
  for $j \leftarrow 1, 2, \ldots, N$ ▷ Can parallelize
    $\vec{z}^0 \leftarrow -\frac{1}{\rho}\vec{\lambda}_i - \vec{x} + \vec{x}_i$
    $t \leftarrow \begin{cases} 1 - 1/\rho\|\vec{z}^0\|_2 & \text{when } \rho\|\vec{z}^0\|_2 \geq 1 \\ 0 & \text{otherwise} \end{cases}$
    $\vec{z}_j \leftarrow t\vec{z}^0$ ▷ $\vec{z}$ update
    $\vec{\lambda}_j \leftarrow \vec{\lambda}_j + \rho(\vec{z}_i + \vec{x} - \vec{x}_i)$ ▷ Dual update

The two examples above show the typical ADMM strategy, in which a difficult nonlinear problem is split into two subproblems that can be carried out in closed form or via more efficient operations. The art of posing a problem in terms of $\vec{x}$ and $\vec{z}$ to get these savings requires practice and careful study of individual problems.

The parameter $\rho > 0$ often does not affect whether or not ADMM will eventually converge, but an intelligent choice of ρ can help this technique converge much faster. Some experimentation can be required here, or ρ can be adjusted from iteration to iteration depending on whether the primal or dual variables are converging more quickly [121]. In some cases, ADMM provably converges faster when $\rho \to \infty$ as the iterations proceed [99].

11.4 GLOBAL OPTIMIZATION

Nonlinear least squares, IRLS, and alternation are lightweight approaches for nonlinear objectives that can be optimized quickly after simplification. On the other side of the spectrum, however, some minimization problems not only do not readily admit fast specialized optimizations but also represent failure modes for Newton's method and other generic solvers. In particular, convergence results for Newton's method and others based on the Taylor approximation assume that we have a strong initial guess of the minimum that we wish to refine. When we lack such an initial guess or a simplifying assumption like convexity, we must solve a *global optimization* problem searching over the entire space of possible inputs to the objective.

As discussed briefly in §8.2, global optimization is a challenging, nearly ill-posed problem. For example, in the unconstrained case it is difficult to know whether $\vec{x}^*$ yields the minimum possible $f(\vec{x})$ *anywhere*, since this is a statement over an infinitude of points $\vec{x}$. Hence, global optimization methods use one or more strategies to improve the odds of finding a minimum:

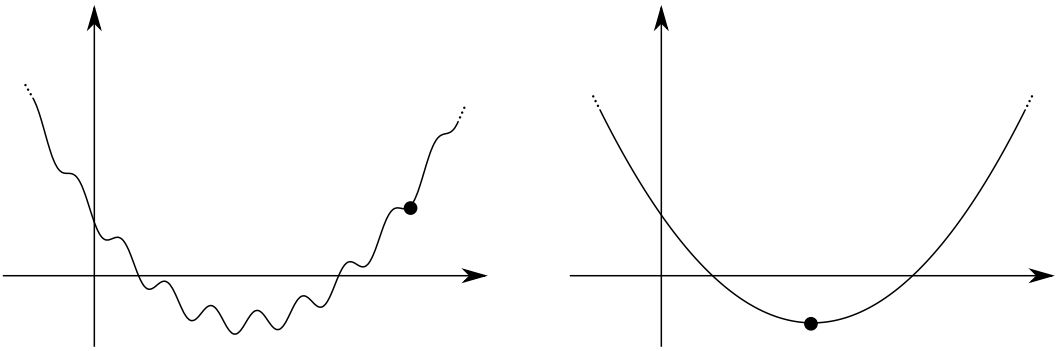


FIGURE 11.4 Newton's method can get caught in any number of local minima in the function on the left; smoothing this function, however, can generate a stronger initial guess of the global optimum.

- Initially approximate the objective $f(\vec{x})$ with an easier function to minimize to get a better starting point for the original problem
- Sample the space of possible inputs $\vec{x}$ to get a better idea of the behavior of f over a large domain

These and other strategies are *heuristic*, meaning that they usually cannot be used to guarantee that the output of such a minimization is globally optimal. In this section, we mention a few common techniques for global optimization as pointers to more specialized literature.

11.4.1 Graduated Optimization

Consider the optimization objective illustrated in Figure 11.4. Locally this objective wiggles up and down, but at a larger scale, a more global pattern emerges. Newton's method seeks *any* critical point of $f(x)$ and easily can get caught in one of its local minima. To avoid this suboptimal output, we might attempt to minimize a smoothed version of $f(x)$ to generate an initial guess for the minimum of the more involved optimization problem.

Graduated optimization techniques solve progressively harder optimization problems with the hope that the coarse initial iterations will generate better initial guesses for the more accurate but sensitive later steps. In particular, suppose we wish to minimize some function $f(\vec{x})$ over $\vec{x} \in \mathbb{R}^n$ with many local optima as in Figure 11.4. Graduated methods generate a sequence of functions $f_1(\vec{x}), f_2(\vec{x}), \ldots, f_k(\vec{x})$ with $f_k(\vec{x}) = f(\vec{x})$, using critical points of f_i as initial guesses for minima of f_{i+1}.

Example 11.10 (Image alignment)**.** A common task making use of graduated optimization is photograph alignment as introduced in §3.1.4. Consider the images in Figure 11.5. Aligning the original two images can be challenging because they have lots of high-frequency detail; for instance, the stones on the wall all look similar and easily could be misidentified. By *blurring* the input images, a better initial guess of the alignment can be obtained, because high-frequency details are suppressed.

The art of graduated optimization lies in finding an appropriate sequence of f_i's to help reach a global optimum. In signal and image processing as in Example 11.10, a typical



Original Blurred

FIGURE 11.5 The photos on the left can be hard to align using automatic methods because they have lots of high-frequency detail that can obscure larger alignment patterns; by blurring the photos we can align larger features before refining the alignment using texture and other detail.

approach is to use the same optimization objective in each iteration but blur the underlying data to reveal larger-scale patterns. *Scale space* methods like [76] instead blur the objective itself, for instance by defining f_i to be $f(\vec{x}) * g_{\sigma_i}(\vec{x})$, the result of blurring $f(\vec{x})$ using a Gaussian of width σ_i, with $\sigma_i \to 0$ as $i \to \infty$.

A related set of optimization techniques known as *homotopy continuation methods* makes continuous changes to the optimization objective by leveraging intuition from topology about homotopic functions:

Definition 11.1 (Homotopic functions)**.** Two continuous functions $f(\vec{x})$ and $g(\vec{x})$ are *homotopic* if there exists continuous function $H(\vec{x}, s)$ with:

$$H(\vec{x}, 0) = f(\vec{x})$$
$$H(\vec{x}, 1) = g(\vec{x})$$

The idea of homotopy is illustrated in Figure 11.6.

Similar to graduated methods, homotopy optimizations minimize $f(\vec{x})$ by defining a new function $H(\vec{x}, s)$ where $H(\vec{x}, 0)$ is easy to optimize and $H(\vec{x}, 1) = f(\vec{x})$. Taking $\vec{x}_0^*$ to be the minimum of $H(\vec{x}, 0)$ with respect to $\vec{x}$, basic homotopy methods incrementally increase s, each time updating to a new $\vec{x}_s^*$. Assuming H is continuous, we expect the minimum $\vec{x}_s^*$ to trace a continuous path in $\mathbb{R}^n$ as s increases; hence, the solve for each $\vec{x}_s^*$ after increasing s differentially has a strong initial guess from the previous iteration.

Example 11.11 (Homotopy methods, [42])**.** Homotopy methods also apply to root-finding. As a small example, suppose we wish to find points x satisfying $\arctan(x) = 0$. Applying the formula from §7.1.4, Newton's method for finding such a root iterates

$$x_{k+1} = x_k - (1 + x_k^2)\arctan(x)$$

If we provide an initial guess $x_0 = 4$, however, this iteration diverges. Instead, we can define a homotopy function as

$$H(x, s) \equiv \arctan(x) + (s - 1)\arctan(4)$$

We know $H(x, 0) = \arctan(x) - \arctan(4)$ has a root at the initial guess $x_0 = 4$. Stepping

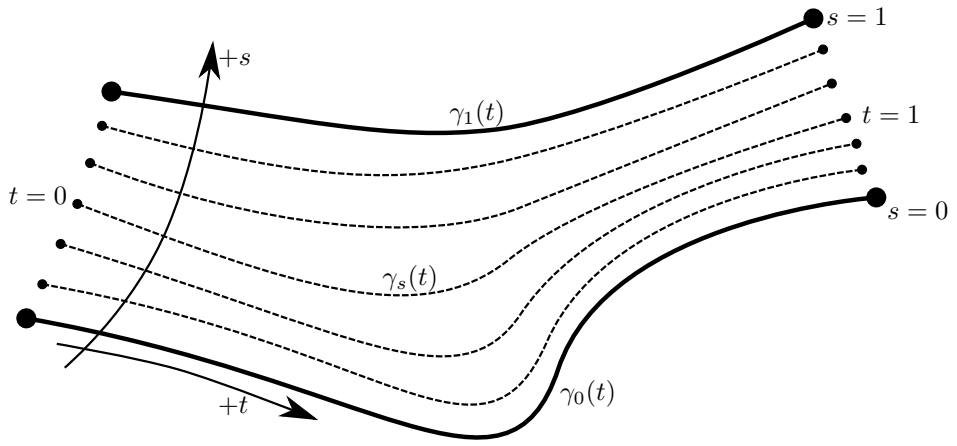


FIGURE 11.6 The curves $\gamma_0(t)$ and $\gamma_1(t)$ are homotopic because there exists a continuously-varying set of curves $\gamma_s(t)$ for $s \in [0, 1]$ coinciding with γ_0 at $s = 0$ and γ_1 at $s = 1$.

s by increments of $1/10$ from 0 to 1, each time minimizing $H(x, s_i)$ with initial guess x^*_{i-1} via Newton's method yields a sequence of convergent problems reaching $x^* = 0$.

More generally, we can think of a *solution path* as a curve of points $(\vec{x}(t), s(t))$ such that $s(0) = 0$, $s(1) = 1$, and at each time t, $\vec{x}(t)$ is a local minimizer of $H(\vec{x}, s(t))$ over $\vec{x}$. Our initial description of homotopy optimization would take $s(t) = t$, but now we can allow $s(t)$ to be non-monotonic as a function of t as long as it eventually reaches $s = 1$. Advanced homotopy continuation methods view $(\vec{x}(t), s(t))$ as a curve satisfying certain ordinary differential equations, which you will derive in Exercise 11.3; these equations can be solved using the techniques we will define in Chapter 14.

11.4.2 Randomized Global Optimization

When smoothing the objective function is impractical or fails to remove local minima from $f(\vec{x})$, it makes sense to sample the space of possible inputs $\vec{x}$ to get some idea of the energy landscape. Newton's method, gradient descent, and others all have strong dependence on the initial guess of the location of the minimum, so trying more than one starting point increases the chances of success.

If the objective f is sufficiently noisy, we may wish to remove dependence on differential estimates altogether. Without gradients, we do not know which directions locally point downhill, but via sampling we can find such patterns on a larger scale. Heuristics for global optimization at this scale commonly draw inspiration from the natural world and the idea of *swarm intelligence*, that complex natural processes can arise from individual actors following simple rules, often in the presence of stochasticity, or randomness. For instance, optimization routines have been designed to mimic ant colonies transporting food [24], thermodynamic energy in "annealing" processes [68], and evolution of DNA and genetic material [82]. These methods usually are considered heuristics without convergence guarantees but can help guide a large-scale search for optima.

As one example of a method well-tuned to continuous problems, we consider the *particle swarm* method introduced in [67] as an optimization technique inspired by social behavior

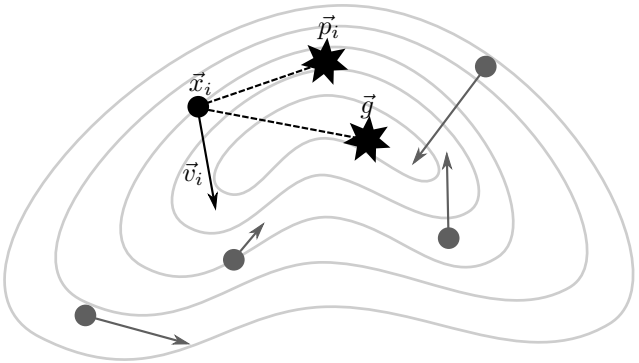


FIGURE 11.7 The particle swarm navigates the landscape of $f(\vec{x})$ by maintaining positions and velocities for a set of potential minima $\vec{x}_i$; each $\vec{x}_i$ is attracted to the position $\vec{p}_i$ at which it has observed the smallest value of $f(\vec{x}_i)$ as well as to the minimum $\vec{g}$ observed thus far by any particle.

```
function PARTICLE-SWARM(f(x⃗), k, α, β, x⃗_min, x⃗_max, v⃗_min, v⃗_max)
  f_min ← ∞
  for i ← 1, 2, ..., k
    x⃗_i ← RANDOM-POSITION(x⃗_min, x⃗_max)        ▷ Initialize positions randomly
    v⃗_i ← RANDOM-VELOCITY(v⃗_min, v⃗_max)        ▷ Initialize velocities randomly
    f_i ← f(x⃗_i)                                ▷ Evaluate f
    p⃗_i ← x⃗_i                                   ▷ Current particle optimum
    if f_i < f_min then                          ▷ Check if it is global optimum
      f_min ← f_i                                ▷ Update optimal value
      g⃗ ← x⃗_i                                   ▷ Set global optimum
  for j ← 1, 2, ...                              ▷ Stop when satisfied with g⃗
    for i ← 1, 2, ..., k
      v⃗_i ← v⃗_i + α(p⃗_i − x⃗_i) + β(g⃗ − x⃗_i)      ▷ Update velocity
      x⃗_i ← x⃗_i + v⃗_i                           ▷ Update position
    for i ← 1, 2, ..., k
      if f(x⃗_i) < f_i then                       ▷ Better minimum for particle i
        p⃗_i ← x⃗_i                               ▷ Update particle optimum
        f_i ← f(x⃗_i)                            ▷ Store objective value
        if f_i < f_min then                      ▷ Check if it is a global optimum
          f_min ← f_i                            ▷ Update optimal value
          g⃗ ← x⃗_i                               ▷ Global optimum
```

FIGURE 11.8 The particle swarm optimization algorithm attempts to minimize $f(\vec{x})$ by simulating a collection of particles $\vec{x}_1, \ldots, \vec{x}_k$ moving in the space of potential inputs $\vec{x}$.

in bird flocks and fish schools. Many variations of this technique have been proposed, but we explore one of the original versions introduced in [33].

Suppose we have a set of candidate minima $\vec{x}_1, \ldots, \vec{x}_k$. We will think of these points as particles moving around the possible space of $\vec{x}$ values, and hence they will also be assigned velocities $\vec{v}_1, \ldots, \vec{v}_k$. The particle swarm method maintains a few additional variables:

- $\vec{p}_1, \ldots, \vec{p}_k$, the position over all iterations so far of the lowest value $f(\vec{p}_i)$ observed by each particle i
- The position $\vec{g} \in \{\vec{p}_1, \ldots, \vec{p}_k\}$ with the smallest objective value; this position is the *globally* best solution observed so far.

This notation is illustrated in Figure 11.7.

In each iteration of particle swarm optimization, the velocities of the particles are updated to guide them toward likely minima. In particular, each particle is attracted to its own best observed minimum as well as to the global best position so far:

$$\vec{v}_i \leftarrow \vec{v}_i + \alpha(\vec{p}_i - \vec{x}_i) + \beta(\vec{g} - \vec{x}_i)$$

The parameters $\alpha, \beta > 0$ determine the amount of force felt from $\vec{x}_i$ to these two positions; larger α, β values will push particles toward minima faster at the cost of more limited exploration of the space of possible minima. Once velocities have been updated, the particles move along their velocity vectors:

$$\vec{x}_i \leftarrow \vec{x}_i + \vec{v}_i$$

Then, the process repeats. This algorithm is not guaranteed to converge, but the algorithm can be terminated at any point, yielding $\vec{g}$ as the best observed minimum. The final method is documented in Figure 11.8.

11.5 ONLINE OPTIMIZATION

We briefly consider a class of optimization problems from machine learning, game theory, and related fields in which the objective itself is allowed to change from iteration to iteration. These problems, known as *online optimization* problems, reflect a world in which evolving input parameters, priorities, and desired outcomes can make the output of an optimization irrelevant soon after it is generated. Hence, techniques in this domain must adaptively react to the changing objective in the presence of noise. Our discussion will introduce a few basic ideas from [102]; we refer the reader to that survey article for a more detailed treatment.

Example 11.12 (Stock market)**.** Suppose we run a financial institution and wish to maintain an optimal portfolio of investments. On the morning of day t, in a highly-simplified model we might choose how much of each stock $1, \ldots, n$ to buy, represented by a vector $\vec{x}_t \in (\mathbb{R}^+)^n$. At the end of the day, based on fluctuations of the market we will know a function f_t so that $f_t(\vec{x})$ gives us our total profit or loss based on the decision $\vec{x}$ made in the morning. The function f_t can be different every day, so we attempt to design a policy that predicts the objective function and/or its optimal point every day.

A model considered in basic online optimization texts related to our example is that of *online convex optimization.* In the unconstrained case, online convex optimization algorithms proceed as follows:

for $t = 1, 2, \ldots$ ▷ At each time t
- ▷ Predict $\vec{x}_t \in U$
- ▷ Receive loss function $f_t : U \to \mathbb{R}$
- ▷ Suffer loss $f_t(\vec{x}_t)$

We will assume the f_t's are convex and that $U \subseteq \mathbb{R}^n$ is a convex set. There are a few features of this setup worth highlighting:

- To stay consistent with our discussion of optimization in previous chapters, we phrase the problem as *minimizing loss* rather than e.g. maximizing profits.
- The optimization objective can change at each time t, and we do not get to know the objective f_t before choosing $\vec{x}_t$. In the stock market example, this feature reflects the fact that we do not know the price of a stock on day t until the day is over, and we must decide how much to buy before getting to that point.
- The online convex optimization algorithm can choose to store $f_1, \ldots, f_{t-1}$ to inform its decision of $\vec{x}_t$. For stock investment, we can use the stock prices on previous days to predict them for the future.

Since online convex optimization algorithms do not know f_t before predicting $\vec{x}_t$, we cannot expect them to perform perfectly. An "adversarial" client might wait for $\vec{x}_t$ and purposefully choose a loss function f_t to make $\vec{x}_t$ look bad! For this reason, metrics like *cumulative loss* $\sum_{t=1}^{T} f_t(\vec{x}_t)$ are unfair measures for the quality of an online optimization method at time T. In some sense, we must lower our standards for success.

One model for online convex optimization is minimization of *regret*, which compares performance to that of a fixed expert benefiting from hindsight:

Definition 11.2 (Regret)**.** The *regret* of an online optimization algorithm at time T over a set U is given by

$$R_T \equiv \max_{\vec{u} \in U} \left[\sum_{t=1}^{T} (f_t(\vec{x}_t) - f_t(\vec{u})) \right].$$

The regret R_T measures the difference between how well our algorithm has performed over time—as measured by summing $f_t(\vec{x}_t)$ over t—and the performance of any constant point $\vec{u}$ that must remain the same over all t. For the stock example, regret compares the profits lost by using our algorithm and the loss of using any single stock portfolio over all time. Ideally, the ratio R_T/T measuring average regret over time should decrease as $T \to \infty$.

The most obvious approach to online optimization is the "follow the leader" (FTL) strategy, which chooses $\vec{x}_t$ based on how it would have performed at times $1, \ldots, t-1$:

$$\textbf{Follow the leader: } \vec{x}_t \equiv \arg\min_{\vec{x} \in U} \sum_{s=1}^{t-1} f_s(\vec{x})$$

FTL is a reasonable heuristic if we assume past performance has some bearing on future results. After all, if we do not know f_t we might as well hope that it is similar to the objectives $f_1, \ldots, f_{t-1}$ we have observed in the past.

For many classes of functions f_t, FTL is an effective approach that makes increasingly well-informed choices of $\vec{x}_t$ as t progresses. It can experience some serious drawbacks, however, as illustrated in the following example:

Example 11.13 (Failure of FTL, [102] §2.2)**.** Suppose $U = [0, 1]$ and we generate a sequence of functions as follows:

$$f_t(x) = \begin{cases} -x/2 & \text{if } t = 1 \\ x & \text{if } t \text{ is even} \\ -x & \text{otherwise} \end{cases}$$

FTL minimizes the sum over all previous objective functions, giving the following series of outputs:

$$\begin{aligned} \mathbf{t = 1}: &\quad x \text{ arbitrary } \in [0, 1] \\ \mathbf{t = 2}: &\quad x_2 = \arg\min_{x \in [0,1]} -x/2 = 1 \\ \mathbf{t = 3}: &\quad x_3 = \arg\min_{x \in [0,1]} x/2 = 0 \\ \mathbf{t = 4}: &\quad x_4 = \arg\min_{x \in [0,1]} -x/2 = 1 \\ \mathbf{t = 5}: &\quad x_5 = \arg\min_{x \in [0,1]} x/2 = 0 \\ \vdots &\qquad\quad \vdots \end{aligned}$$

From the above calculation, we find that in every iteration except $t = 1$, FTL incurs loss 1, while fixing $x = 0$ for all time would incur zero loss. Hence, for this example FTL has regret growing proportional to t.

This example illustrates the type of analysis and reasoning typically needed to design online learning methods. To bound regret, we must consider the *worst*-possible adversary, who generates functions f_t specifically designed to take advantage of the weaknesses of a given technique.

FTL failed because it was too strongly sensitive to the fluctuations of f_t from iteration to iteration. To resolve this issue, we can take inspiration from Tikhonov regularization (§3.1.3), L^1 regularization (§9.4.1), and other methods that dampen the output of numerical methods by adding an energy term punishing irregular or large output vectors. To do so, we define the "follow the regularized leader" (FTRL) strategy:

$$\textbf{Follow the regularized leader: } \vec{x}_t \equiv \arg\min_{\vec{x} \in U} \left[r(\vec{x}) + \sum_{s=1}^{t-1} f_s(\vec{x}) \right]$$

Here, $r(\vec{x})$ is a convex regularization function, such as $\|\vec{x}\|_2^2$ (Tikhonov regularization), $\|\vec{x}\|_1$ (L^1 regularization), or $\sum_i x_i \log x_i$ when U includes only $\vec{x} \geq \vec{0}$ (entropic regularization).

Just as regularization improves the conditioning of a linear problem when it is close to singular, in this case the change from FTL to FTRL avoids fluctuation issues illustrated in Example 11.13. For instance, suppose $r(\vec{x})$ is *strongly convex* as defined below for differentiable r:

Definition 11.3 (Strongly convex)**.** A differentiable regularizer $r(\vec{x})$ is *σ-strongly-convex* with respect to a norm $\|\cdot\|$ if for any $\vec{x}, \vec{y}$ the following relationship holds:

$$(\nabla r(\vec{x}) - \nabla r(\vec{y})) \cdot (\vec{x} - \vec{y}) \geq \sigma \|\vec{x} - \vec{y}\|_2^2$$

Intuitively, a strongly convex regularizer not only is bowl-shaped but has a lower bound for the curvature of that bowl. Then, we we can prove the following statement:

Proposition 11.1 ([102], Theorem 2.11)**.** Assume $r(\vec{x})$ is σ-strongly-convex and that each f_t is convex and L-Lipschitz (see §7.1.1). Then, the regret is bounded as follows:

$$R_T \leq \left[\max_{\vec{u} \in U} r(\vec{u}) \right] - \left[\min_{\vec{v} \in U} r(\vec{v}) \right] + \frac{TL^2}{\sigma}.$$

The proof of this proposition uses techniques well within the scope of this book but due to its length is omitted from our discussion.

Proposition 11.1 can be somewhat hard to interpret, but it is a strong result about the effectiveness of the FTRL technique given an appropriate choice of r. In particular, the max and min terms as well as σ are properties of $r(\vec{x})$ that should guide which regularizer to use for a particular problem. The value σ contributes to both terms in competing ways:

- The difference of the maximum and minimum values of r is its range of possible outputs. Increasing σ has the potential to increase this difference since it is bounded below by a "steeper" bowl. So, minimizing this term in our regret bound prefers *small* sigma.
- Minimizing TL^2/σ prefers *large* sigma.

Practically speaking, we can decide what range of T we care about and choose a regularizer accordingly:

Example 11.14 (FTRL choice of regularizers)**.** Consider the regularizer $r_\sigma(\vec{x}) \equiv \frac{1}{2}\sigma\|\vec{x}\|_2^2$. It has gradient $\nabla r_\sigma(\vec{x}) = \sigma\vec{x}$, so by direct application of Definition 11.3 it is σ-strongly-convex. Suppose $U = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\|_2 \leq 1\}$ and that we expect to run our optimization for T time steps. If we take $\sigma = \sqrt{T}$, then the regret bound from Proposition 11.1 shows:

$$R_T \leq (1 + L^2)\sqrt{T}$$

For large T, this value is small relative to T, compared to the linear growth for FTL in Example 11.13.

Online optimization is a rich area of research that continues to be explored. Beyond FTRL, we can define algorithms with better or more usable regret bounds, especially if we know more about the class of functions f_t we expect to observe. FTRL also has the drawback that it has to solve a potentially complex optimization problem at each iteration, which may not be practical for systems that have to make decisions quickly. Surprisingly, even easy-to-solve linearizations can behave fairly well for convex objectives, as illustrated in problem 11.11. Popular online optimization techniques like [31] have been applied to a variety of learning problems in the presence of huge amounts of noisy data.

11.6 EXERCISES

11.1 An alternative derivation of the Gauss-Newton algorithm shows that it can be thought of as an approximation of Newton's method for unconstrained optimization.

(a) Write an expression for the Hessian of $E_{\text{NLS}}(\vec{x})$ (defined in §11.1) in terms of the derivatives of the f_i's.

(b) Show that the Gauss-Newton algorithm on E_{NLS} is equivalent to Newton's method (§8.4.2) after removing second derivative terms from the Hessian.

(c) When is such an approximation of the Hessian reasonable?

11.2 For a fixed parameter $\delta > 0$, the *Huber loss function* $L_\delta(x)$ is defined as:

$$L_\delta(x) \equiv \begin{cases} x^2/2, & \text{when } |x| \leq \delta \\ \delta(|x| - \delta/2), & \text{otherwise.} \end{cases}$$

This function "softens" the non-differentiable singularity of $|x|$ at $x = 0$.

(a) Illustrate the effect of choosing different values of δ on the shape of $L_\delta(x)$.

(b) Recall that we can find an $\vec{x}$ nearly satisfying the overdetermined system $A\vec{x} \approx \vec{b}$ by minimizing $\|A\vec{x} - \vec{b}\|_2$ (least squares) or $\|A\vec{x} - \vec{b}\|_1$ (compressive sensing). Propose a similar optimization compromising between these two methods using L_δ.

(c) Propose an IRLS algorithm for optimizing your objective from part 11.2b.

(d) Propose an ADMM algorithm for optimizing your objective from part 11.2b. *Hint:* Introduce a variable $\vec{z} = A\vec{x} - \vec{b}$.

11.3[DH] In §11.4.1, we introduced homotopy continuation methods for optimization. These methods begin by minimizing a simple objective $H(\vec{x}, 0) = f_0(\vec{x})$ and proceed by solving continuously-modified objectives until a minimum of $H(\vec{x}, 1) = f(\vec{x})$ (the original problem objective) is found.

Suppose that instead of the time function $s(t) = t$ used in Example 11.11, we let $s(t)$ be an arbitrary function of t such that $s(0) = 0$. We will assume that t can take any nonnegative value, and we only require that $s(t)$ eventually reaches 1.

(a) What relationship does $H(\vec{x}(t), s(t))$ satisfy for all $t \geq 0$ for points $(\vec{x}(t), s(t))$ on the solution path?

(b) Differentiate this equation with respect to t. Write one side as the product of two vectors.

(c) What is the geometric interpretation of the vector $\vec{g}(t) \equiv \left(\nabla\vec{x}(t), \frac{d}{dt}s(t)\right)^\top$ in terms of the solution path?

(d) We will impose the restriction that $||\vec{g}(t)||_2^2 = 1\ \forall t$, i.e. that $\vec{g}(t)$ is unit length. What is the geometric interpretation of t, again in terms of the solution path?

(e) Given the initial data $(\vec{x}(0), 0)$, as well as $\vec{g}(t)$, write down an ordinary differential equation (ODE) whose solution is a solution path for $t > 0$. As long as we can evaluate $\vec{x}(t)$, $s(t)$, and their derivatives, numerical ODE solvers can now give us the solution path to our homotopy continuation optimization. This provides a connection between topology, optimization, and differential equations.

11.4 ("Least absolute deviations") Instead of solving least-squares, to take advantage of methods from compressive sensing we might wish to minimize $\|A\vec{x} - \vec{b}\|_1$ with $\vec{x}$ unconstrained. Propose an ADMM-style splitting of this optimization and give the alternating steps of the optimization technique in this case.

11.5[DH] Suppose we have two convex sets $S, T \subseteq \mathbb{R}^n$. The *alternating projection* method discussed in [9] and elsewhere is used to find a point $\vec{x} \in S \cap T$. For any initial guess $\vec{x}_0$, alternating projection performs the iteration

$$\vec{x}_{k+1} = \mathcal{P}_S\left(\mathcal{P}_T\left(\vec{x}_k\right)\right),$$

where $\mathcal{P}_S$ and $\mathcal{P}_T$ are operators that project onto the nearest point in S or T, respectively. As long as $S \cap T \neq \emptyset$, this iterative procedure is guaranteed to converge to an $\vec{x} \in S \cap T$, though this convergence may be impractically slow [21]. Instead of this algorithm, we will consider finding an intersection point of convex sets using ADMM.

(a) Propose an unconstrained optimization problem whose solution is a point $\vec{x} \in S \cap T$, assuming $S \cap T \neq \emptyset$.
Hint: Use indicator functions.

(b) Write this problem in a form that is amenable to ADMM, using $\vec{x}$ and $\vec{z}$ as your variables.

(c) Explicitly write the ADMM iterations for updating $\vec{x}$, $\vec{z}$, and dual variables $\vec{w}$.
Hint: Your expressions need to use $\mathcal{P}_S$ and $\mathcal{P}_T$.

11.6[DH] Another popular technique for global optimization is *simulated annealing* [68], a method motivated by ideas from statistical physics. The term *annealing* refers to the process in metallurgy where a metal is heated and then cooled so its constituent particles arrange in a minimum energy state. In this thermodynamic process, atoms may move considerably at higher temperatures but become restricted in motion as the temperature cools. Borrowing from this analogy, in the context of global optimization, one may imagine letting a potential solution change considerably early on in a search to explore the space of outputs, then become fixed as the number of iterations gets large. Pseudocode for the simulated annealing algorithm is provided in the following box.

```
function SIMULATED-ANNEALING(f(x⃗), x⃗_0)
  T_0 ← High temperature
  T_i ← Cooling schedule, e.g. T_i = αT_{i−1} for some α < 1
  x⃗ ← x⃗_0                          ▷ Current model initialized to the input x⃗_0
  for i ← 1, 2, 3, ...
    y⃗ ← RANDOM-MODEL                          ▷ Random guess of output
    Δf ← f(y⃗) − f(x⃗)                          ▷ Compute change in objective
    if Δf < 0 then                             ▷ Objective improved at y⃗
      x⃗ ← y⃗
    else if UNIFORM(0,1)< e^{−Δf/T_i} then     ▷ True with probability e^{−Δf/T_i}
      x⃗ ← y⃗                                   ▷ Randomly keep suboptimal output
```

Simulated annealing randomly guesses solutions in each iteration. If the new solution is more optimal than the current solution, the algorithm moves forward using the new solution. If the new solution is less optimal, however, the algorithm does not necessarily reject it; instead, it accepts it with exponentially small probability as temperature decreases. The hope of this heuristic is that local minima will be avoided in favor of global minima due to the significant variation in solutions found during the first few iterations, while some form of convergence is still obtained as the iterates generally stabilize at lower temperatures.

Consider the Euclidean traveling salesman problem (TSP): Given a connected graph in $\mathbb{R}^2$, where cities are vertices and roads are edges, we wish to visit each city exactly once while minimizing the sum of the Euclidean distances traveled between cities in the tour. While Euclidean TSP is NP-hard, simulated annealing provides a practical approximation algorithm to solve this problem in terms of global optimization.

(a) Phrase Euclidean TSP as a global optimization problem.

(b) Generating random tours of a graph is computationally expensive. Can you propose a reasonable scheme for generating a random model $\vec{y}$ at each step that does not involve finding a tour?

(c) Implement the simulated annealing solution to Euclidean TSP and explore the trade-off between solution quality and runtime when the initial temperature T_0 is changed. Also, explore the results of different cooling schedules, either by varying α in the example T_i or by presenting your own cooling schedule.

(d) Choose another global optimization algorithm and consider how to use it to solve Euclidean TSP. Can you make any arguments as to how it would compare to simulated annealing?

11.7[SC] Recall the setup from problem 9.7, in which you wish to design a slow-dissolving medicinal capsule shaped as a cylinder with hemispherical ends. To ensure that the capsule dissolves slowly, we need to minimize its surface area with constraints that its length ℓ is greater than some constant $\ell_{\min}$; the entire capsule can be no longer than some length C.

(a) Suppose you were unhappy with the results of the optimization you proposed in problem 9.7 and want to ensure that the entire volume of the capsule is at least V. Explain why the resulting problem cannot be solved using geometric programming methods.

(b) Propose an alternating optimization method for this problem. Is it necessary to solve a geometric program for either alternation?

11.8 The mean shift algorithm, originally proposed in [25], is an iterative clustering technique appearing in literature on nonparametric machine learning and image processing. Given a set of n data points $\vec{x}_i \in \mathbb{R}^d$, the algorithm groups points together based on their closest maxima of a smoothed density function approximating the distribution of data points.

(a) Take $k(x) : \mathbb{R} \to \mathbb{R}^+$ to be a nonnegative function. For a fixed bandwidth parameter $h > 0$, define the *kernel density estimator* $\hat{f}(\vec{x})$ to be

$$\hat{f}_k(\vec{x}) \equiv \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{\vec{x} - \vec{x}_i}{h} \right\|_2^2 \right)$$

If $k(x)$ is peaked at $x = 0$, explain how $\hat{f}_k(\vec{x})$ encodes the density of data points $\vec{x}_i$. What is the effect of increasing the parameter h?
Note: The constant $c_{k,d}$ is chosen so that $\int_{\mathbb{R}^d} \hat{f}(\vec{x})\, d\vec{x} = 1$. Choosing $k(x) \equiv e^{-x/2}$ makes $\hat{f}$ a sum of Gaussians.

(b) Define $g(x) \equiv -k'(x)$ and take $m(\vec{x})$ to be the *mean shift* vector given by

$$m(\vec{x}) \equiv \frac{\sum_i \vec{x}_i g\left(\left\| \frac{\vec{x} - \vec{x}_i}{h} \right\|_2^2 \right)}{\sum_i g\left(\left\| \frac{\vec{x} - \vec{x}_i}{h} \right\|_2^2 \right)} - \vec{x}.$$

Show that $\nabla \hat{f}_k(\vec{x})$ can be factored as follows:

$$\nabla \hat{f}_k(\vec{x}) = \frac{\alpha}{h^2} \cdot \hat{f}_g(\vec{x}) \cdot m(\vec{x}),$$

for some constant α.

(c) Suppose $\vec{y}_0$ is a guess of the location of a peak of $\hat{f}_k$. Using your answer from part 11.8b, motivate the *mean shift* algorithm for finding a peak of $\hat{f}_k(\vec{x})$, which iterates the formula

$$\vec{y}_{k+1} \equiv \frac{\sum_i \vec{x}_i g\left(\left\|\frac{\vec{y}-\vec{x}_i}{h}\right\|_2^2\right)}{\sum_i g\left(\left\|\frac{\vec{y}-\vec{x}_i}{h}\right\|_2^2\right)}.$$

Note: This algorithm is guaranteed to converge under mild conditions on k. Mean shift clustering runs this method to convergence starting from $\vec{y}_0 = \vec{x}_i$ for each i in parallel; $\vec{x}_i$ and $\vec{x}_j$ are assigned to the same cluster if mean shift iteration yields the same output (within some tolerance) for starting points $\vec{y}_0 = \vec{x}_i$ and $\vec{y}_0 = \vec{x}_j$.

(d) Suppose we represent a grayscale image as a set of pairs $(\vec{p}_i, q_i)$, where $\vec{p}_i$ is the center of pixel i (typically laid out on a grid), and $q_i \in [0, 1]$ is the intensity of pixel i. The *bilateral filter* [114] for blurring images while preserving their sharp edges is given by:

$$\hat{q}_i \equiv \frac{\sum_j q_j k_1(\|\vec{p}_j - \vec{p}_i\|_2) k_2(|q_j - q_i|)}{\sum_j k_1(\|\vec{p}_j - \vec{p}_i\|_2) k_2(|q_j - q_i|)},$$

where k_1, k_2 are Gaussian kernels given by $k_i(x) \equiv e^{-a_i x^2}$. Fast algorithms have been developed in the computer graphics community for evaluating the bilateral filter and its variants [92].

Propose an algorithm for clustering the pixels in an image using iterated calls to a modified version of the bilateral filter; the resulting method is called the "local mode filter" [119, 91].

11.9 The *iterative shrinkage-thresholding algorithm* (ISTA) is another technique relevant to large-scale optimization applicable to common objectives from machine learning. Extensions such as [11] have led to renewed interest in this technique. We follow development of [18].

(a) Show that the iteration from gradient descent

$$\vec{x}_{k+1} = \vec{x}_k - \alpha \nabla f(\vec{x}_k)$$

can be rewritten in *proximal form* as

$$\vec{x}_{k+1} = \arg\min_{\vec{x}} \left[f(\vec{x}_k) + \nabla f(\vec{x}_k)^\top (\vec{x} - \vec{x}_k) + \frac{1}{2\alpha} \|\vec{x} - \vec{x}_k\|_2^2 \right].$$

(b) Suppose we wish to minimize a sum $f(\vec{x}) + g(\vec{x})$. Based on the previous part, ISTA attempts to combine exact optimization for g with gradient descent on f:

$$\vec{x}_{k+1} \equiv \arg\min_{\vec{x}} \left[f(\vec{x}_k) + \nabla f(\vec{x}_k)^\top (\vec{x} - \vec{x}_k) + \frac{1}{2\alpha} \|\vec{x} - \vec{x}_k\|_2^2 + g(\vec{x}) \right].$$

Derive the alternative form

$$\vec{x}_{k+1} = \arg\min_{\vec{x}} \left[g(\vec{x}) + \frac{1}{2\alpha} \|\vec{x} - (\vec{x}_k - \alpha \nabla f(\vec{x}_k))\|_2^2 \right].$$

(c) Derive a formula for ISTA iterations when $g(\vec{x}) = \lambda\|\vec{x}\|_1$.
Hint: This case reduces to solving a set of single-variable problems.

11.10 Suppose $\mathcal{D}$ is a bounded, convex, and closed domain in $\mathbb{R}^n$ and $f(\vec{x})$ is a convex, differentiable objective function. The *Frank-Wolfe* algorithm for minimizing $f(\vec{x})$ subject to $\vec{x} \in \mathcal{D}$ is as follows [40]:

$$\begin{aligned}\vec{s}_k &\leftarrow \arg\min_{\vec{s}\in\mathcal{D}}[\vec{s}\cdot\nabla f(\vec{x}_{k-1})]\\ \gamma_k &\leftarrow \frac{2}{k+2}\\ \vec{x}_k &\leftarrow (1-\gamma_k)\vec{x}_{k-1}+\gamma_k\vec{s}_k\end{aligned}$$

A starting point $\vec{x}_0 \in \mathcal{D}$ must be provided. This algorithm has gained renewed attention for large-scale optimization in machine learning in the presence of sparsity and other specialized structure [61].

(a) Argue that $\vec{s}_k$ minimizes a linearized version of f subject to the constraints. Also, show that if $\mathcal{D} = \{\vec{x} : A\vec{x} \le \vec{b}\}$ for fixed $A \in \mathbb{R}^{m\times n}$ and $\vec{b} \in \mathbb{R}^m$, then each iteration of the Frank-Wolfe algorithm solves a linear program.

(b) Show that $\vec{x}_k \in \mathcal{D}$ for all $k > 0$.

(c) Assume $\nabla f(\vec{x})$ is L-Lipschitz on $\mathcal{D}$, meaning $\|\nabla f(\vec{x}) - \nabla f(\vec{y})\|_2 \le L\|\vec{x}-\vec{y}\|_2$, for all $\vec{x}, \vec{y} \in \mathcal{D}$. Derive the bound (proposed in [83]):

$$|f(\vec{y}) - f(\vec{x}) - (\vec{y}-\vec{x})\cdot\nabla f(\vec{x})| \le \frac{L}{2}\|\vec{y}-\vec{x}\|_2^2.$$

Hint: By the Fundamental Theorem of Calculus, $f(\vec{y}) = f(\vec{x}) + \int_0^1 (\vec{y}-\vec{x})\cdot\nabla f(\vec{x}+\tau(\vec{y}-\vec{x}))\,d\tau$.

(d) Define the *diameter* of $\mathcal{D}$ to be $d \equiv \max_{\vec{x},\vec{y}\in\mathcal{D}}\|\vec{x}-\vec{y}\|_2$. Furthermore, assume $\nabla f(\vec{x})$ is L-Lipschitz on $\mathcal{D}$. Show that

$$\frac{2}{\gamma^2}(f(\vec{y}) - f(\vec{x}) - (\vec{y}-\vec{x})\cdot\nabla f(\vec{x})) \le d^2 L,$$

for all $\vec{x}, \vec{y}, \vec{s} \in \mathcal{D}$ with $\vec{y} = \vec{x} + \gamma(\vec{s}-\vec{x})$ and $\gamma \in [0,1]$. Conclude that

$$f(\vec{y}) \le f(\vec{x}) + \gamma(\vec{s}-\vec{x})\cdot\nabla f(\vec{x}) + \frac{\gamma^2 d^2 L}{2}.$$

(e) Define the *duality gap* $g(\vec{x}) \equiv \max_{\vec{s}\in\mathcal{D}}(\vec{x}-\vec{s})\cdot\nabla f(\vec{x})$. For the Frank-Wolfe algorithm show

$$f(\vec{x}_k) \le f(\vec{x}_{k-1}) - \gamma g(\vec{x}_{k-1}) + \frac{\gamma_k^2 d^2 L}{2}.$$

(f) Take $\vec{x}^*$ to be the location of the minimum for the optimization problem, and define $h(\vec{x}) \equiv f(\vec{x}) - f(\vec{x}^*)$. Show $g(\vec{x}) \ge h(\vec{x})$, and using the previous part conclude

$$h(\vec{x}_k) \le (1-\gamma_k)h(\vec{x}_{k-1}) + \frac{\gamma_k^2 d^2 L}{2}.$$

(g) Conclude $h(\vec{x}_k) \to 0$ as $k \to \infty$. What does this imply about the Frank-Wolfe algorithm?

11.11 The FTRL algorithm from §11.5 can be expensive when the f_t's are difficult to minimize. In this problem, we derive a linearized alternative with similar performance guarantees.

(a) Suppose we make the following assumptions about an instance of FTRL:
- $U = \{\vec{x} \in \mathbb{R}^n : \|\vec{x}\|_2 \leq 1\}$
- All of the objectives f_t provided to FTRL are of the form $f_t(\vec{x}) = \vec{z}_t \cdot \vec{x}$ for $\|\vec{z}_t\|_2 \leq 1$.
- $r(\vec{x}) \equiv \frac{1}{2}\sigma\|\vec{x}\|_2^2$

Provide an explicit formula for the iterates $\vec{x}_t$ in this case, and specialize the bound from Proposition 11.1 to this case.

(b) We wish to apply the bound from 11.11a to more general f_t's. To do so, suppose we replace FTRL with a linearized objective for $\vec{x}_t$:

$$\vec{x}_t \equiv \arg\min_{\vec{x} \in U} \left[r(\vec{x}) + \sum_{s=1}^{t-1} (f_s(\vec{x}_{t-1}) + \nabla f_s(\vec{x}_{t-1}) \cdot (\vec{x} - \vec{x}_{t-1})) \right].$$

Provide an explicit formula for $\vec{x}_t$ in this case, assuming the same choice of U and r.

(c) Propose a regret bound for the linearized method in 11.11b.
Hint: Apply convexity of the f_t's and the result of 11.11a.

IV

Functions, Derivatives, and Integrals

CHAPTER 12

Interpolation

CONTENTS

12.1 Interpolation in a Single Variable 250
12.1.1 Polynomial Interpolation 250
12.1.2 Alternative Bases 253
12.1.3 Piecewise Interpolation 255
12.2 Multivariable Interpolation 257
12.2.1 Nearest-Neighbor Interpolation 257
12.2.2 Barycentric Interpolation 258
12.2.3 Grid-Based Interpolation 260
12.3 Theory of Interpolation 260
12.3.1 Linear Algebra of Functions 261
12.3.2 Approximation via Piecewise Polynomials 263

SO far we have derived methods for *analyzing* functions f, e.g. finding their minima and roots. Evaluating $f(\vec{x})$ at a particular $\vec{x} \in \mathbb{R}^n$ might be expensive, but a fundamental assumption of the methods we developed in previous chapters is that we can obtain $f(\vec{x})$ when we want it, regardless of $\vec{x}$.

There are many contexts in which this assumption is unrealistic. For instance, if we take a photograph with a digital camera, we receive an $n \times m$ grid of pixel color values sampling the continuum of light coming into the camera lens. That is, we might think of a photograph as a continuous function from image position (x, y) to color (r, g, b), but in reality we only know the image value at nm separated locations on the image plane. Similarly, in machine learning and statistics, often we only are given samples of a function at points where we collected data, and we must interpolate to have values elsewhere; in a medical setting we may monitor a patient's response to different dosages of a drug but must predict what will happen at a dosage we have not tried explicitly.

In these cases, before we can minimize a function, find its roots, or even compute values $f(\vec{x})$ at arbitrary locations $\vec{x}$, we need a model for interpolating $f(\vec{x})$ to all of $\mathbb{R}^n$ (or some subset thereof) given a collection of samples $f(\vec{x}_i)$. Techniques for this *interpolation* problem are inherently approximate, since we do not know the true values of f, so instead we seek for the interpolated function to be smooth and serve as a "reasonable" prediction of function values.

In this chapter, we will assume that the values $f(\vec{x}_i)$ are known with complete certainty; in this case we can think of the problem as extending f to the remainder of the domain without perturbing the value at any of the input locations. To contrast, the *regression* problem considered in §3.1.1 and elsewhere may forgo matching $f(\vec{x}_i)$ completely in favor of making f more smooth.

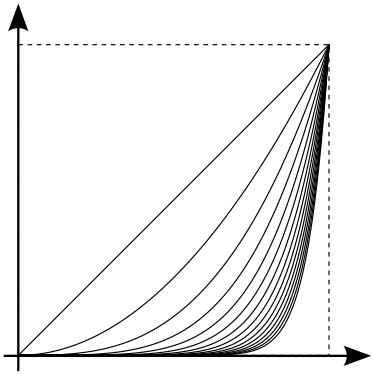


FIGURE 12.1 As k increases, the monomials x^k on $[0, 1]$ begin to look more and more similar. This similarity creates poor conditioning for monomial basis problems like solving the Vandermonde system.

12.1 INTERPOLATION IN A SINGLE VARIABLE

Before considering the general case, we will design methods for interpolating functions of a single variable $f : \mathbb{R} \to \mathbb{R}$. As input, we will take a set of k pairs (x_i, y_i) with the assumption $f(x_i) = y_i$; our job is to predict $f(x)$ for $x \notin \{x_1, \ldots, x_k\}$. Desirable *interpolants* $f(x)$ should be smooth and should interpolate the data points faithfully without adding extra features like spurious local minima and maxima.

We will take inspiration from linear algebra by writing $f(x)$ in a *basis*. That is, the set of all possible functions $f : \mathbb{R} \to \mathbb{R}$ is far too large to work with and includes many functions that are not practical in a computational setting. Thus, we simplify the search space by forcing f to be written as a linear combination of building block basis functions. This formulation is familiar from calculus: The Taylor expansion writes functions in the basis of polynomials, while Fourier series use sine and cosine.

12.1.1 Polynomial Interpolation

Perhaps the most straightforward class of interpolation formulas assumes that $f(x)$ is in $\mathbb{R}[x]$, the set of polynomials. Polynomials are smooth, and we already have explored linear methods for finding a degree $k - 1$ polynomial through k sample points in Chapter 3.

Example 3.3 worked out the details of such an interpolation technique. As a reminder, suppose we wish to find $f(x) \equiv a_0 + a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1}$ through the points $(x_1, y_1), \ldots, (x_k, y_k)$; here our unknowns are the values $a_0, \ldots, a_{k-1}$. Plugging in the expression $y_i = f(x_i)$ for each i shows that the vector $\vec{a}$ satisfies the $k \times k$ *Vandermonde* system:

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{k-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^{k-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{pmatrix}$$

By this construction, degree $k - 1$ polynomial interpolation can be accomplished using a $k \times k$ linear solve for $\vec{a}$ using the linear algorithms in Chapter 2. This method, however, is far from optimal for many applications.

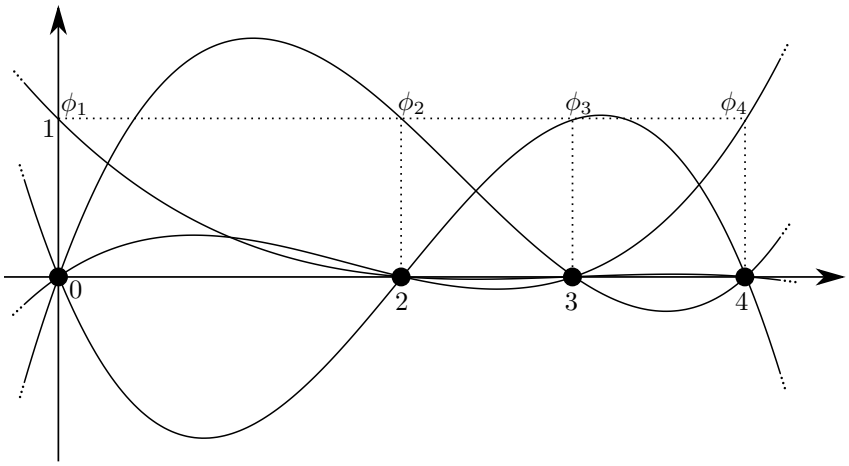


FIGURE 12.2 The Lagrange basis for $x_1 = 0, x_2 = 2, x_3 = 3, x_4 = 4$. Each ϕ_i satisfies $\phi_i(x_i) = 1$ and $\phi_i(x_j) = 0$ for all $i \neq j$.

As mentioned above, one way to think about the space of polynomials is that it can be spanned by a basis of functions. Just as a basis for $\mathbb{R}^n$ is a set of n linearly-independent vectors $\vec{v}_1, \ldots, \vec{v}_n$, in our derivation of the Vandermonde matrix we wrote the space of polynomials of degree $k-1$ as the span of *monomials* $\{1, x, x^2, \ldots, x^{k-1}\}$. Although monomials may be an obvious basis for $\mathbb{R}[x]$, they have limited properties useful for simplifying the polynomial interpolation problem. One way to visualize this issue is to plot the sequence of functions $1, x, x^2, x^3, \ldots$ for $x \in [0, 1]$; in this interval, as shown in Figure 12.1 the functions x^k all start looking similar as k increases. As we know from our consideration of projection problems in Chapter 4, projection onto a set of similar-looking basis vectors can be unstable.

Continuing to apply intuition from linear algebra, we may choose to write polynomials in a basis that is better suited to the problem at hand. This time, recall that we are given k pairs $(x_1, y_1), \ldots, (x_k, y_k)$. We will use these (fixed) points to define the *Lagrange interpolation* basis $\phi_1, \ldots, \phi_k$ by writing:

$$\phi_i(x) \equiv \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}$$

Example 12.1 (Lagrange basis)**.** Suppose $x_1 = 0$, $x_2 = 2$, $x_3 = 3$, and $x_4 = 4$. The Lagrange basis for this set of x_i's is:

$$\begin{aligned}
\phi_1(x) &= \frac{(x-2)(x-3)(x-4)}{-2 \cdot -3 \cdot -4} = \frac{1}{24}(-x^3 + 9x^2 - 26x + 24) \\
\phi_2(x) &= \frac{x(x-3)(x-4)}{2 \cdot (2-3)(2-4)} = \frac{1}{4}(x^3 - 7x^2 + 12x) \\
\phi_3(x) &= \frac{x(x-2)(x-4)}{3 \cdot (3-2) \cdot (3-4)} = \frac{1}{3}(-x^3 + 6x^2 - 8x) \\
\phi_4(x) &= \frac{x(x-2)(x-3)}{4 \cdot (4-2) \cdot (4-3)} = \frac{1}{8}(x^3 - 5x^2 + 6x)
\end{aligned}$$

This basis is shown in Figure 12.2.

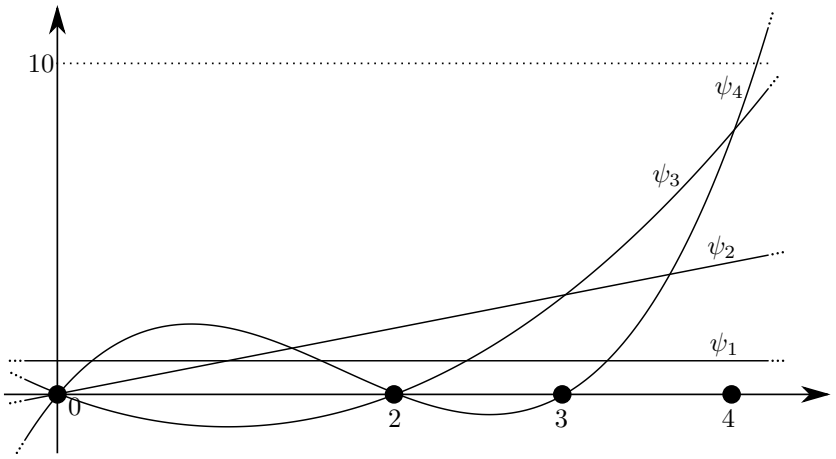


FIGURE 12.3 The Newton basis for $x_1 = 0, x_2 = 2, x_3 = 3, x_4 = 4$. Each ψ_i satisfies $\psi_i(x_j) = 0$ when $j < i$.

As shown in this example, although it is not written in the monomial basis $\{1, x, x^2, \ldots, x^{k-1}\}$, each ϕ_i is still a polynomial of degree $k-1$. Furthermore, the Lagrange basis has the following desirable property:

$$\phi_i(x_\ell) = \begin{cases} 1 & \text{when } \ell = i \\ 0 & \text{otherwise.} \end{cases}$$

Using this formula, finding the unique degree $k-1$ polynomial fitting our (x_i, y_i) pairs is formulaic in the Lagrange basis:

$$f(x) \equiv \sum_i y_i \phi_i(x)$$

To check, if we substitute $x = x_j$ we find:

$$\begin{aligned} f(x_j) &= \sum_i y_i \phi_i(x_j) \\ &= y_j \text{ since } \phi_i(x_j) = 0 \text{ when } i \neq j. \end{aligned}$$

We have shown that in the Lagrange basis we can write a closed formula for $f(x)$ that does not require solving the Vandermonde system. The drawback, however, is that each $\phi_i(x)$ takes $O(k)$ time to evaluate using the formula above, so computing $f(x)$ takes $O(k^2)$ time total; contrastingly, if we find the coefficients a_i from the Vandermonde system explicitly, the evaluation time for interpolation subsequently becomes $O(k)$.

Computation time aside, the Lagrange basis has an additional numerical drawback, in that the denominator is the product of a potentially large number of terms. If the x_i's are close together, then this product may include many terms close to zero; the end result is division by a small number when evaluating $\phi_i(x)$. As we have seen, this operation can create numerical instabilities that we wish to avoid.

A third basis for polynomials of degree $k-1$ that attempts to compromise between the numerical quality of the monomials and the efficiency of the Lagrange basis is the *Newton*

basis, defined as follows:

$$\psi_i(x) = \prod_{j=1}^{i-1}(x - x_j)$$

This expression has no product terms when $i = 1$, so we define $\psi_1(x) \equiv 1$. This way, for all indices i we have that $\psi_i(x)$ is a degree $i - 1$ polynomial.

Example 12.2 (Newton basis)**.** Continuing from Example 12.1, again suppose $x_1 = 0$, $x_2 = 2$, $x_3 = 3$, and $x_4 = 4$. The corresponding Newton basis is:

$$\begin{aligned}
\psi_1(x) &= 1 \\
\psi_2(x) &= x \\
\psi_3(x) &= x(x-2) = x^2 - 2x \\
\psi_4(x) &= x(x-2)(x-3) = x^3 - 5x^2 + 6x
\end{aligned}$$

This basis is illustrated in Figure 12.3.

By definition of ψ_i, $\psi_i(x_\ell) = 0$ for all $\ell < i$. If we wish to write $f(x) = \sum_i c_i\psi_i(x)$ and write out this observation more explicitly, we find:

$$\begin{aligned}
f(x_1) &= c_1\psi_1(x_1) \\
f(x_2) &= c_1\psi_1(x_2) + c_2\psi_2(x_2) \\
f(x_3) &= c_1\psi_1(x_3) + c_2\psi_2(x_3) + c_3\psi_3(x_3) \\
&\vdots \quad \vdots
\end{aligned}$$

In other words, we can solve the following lower triangular system for $\vec{c}$:

$$\begin{pmatrix}
\psi_1(x_1) & 0 & 0 & \cdots & 0 \\
\psi_1(x_2) & \psi_2(x_2) & 0 & \cdots & 0 \\
\psi_1(x_3) & \psi_2(x_3) & \psi_3(x_3) & \cdots & 0 \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\psi_1(x_k) & \psi_2(x_k) & \psi_3(x_k) & \cdots & \psi_k(x_k)
\end{pmatrix}
\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix} =
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}$$

This system can be solved in $O(k^2)$ time using forward substitution, rather than the $O(k^3)$ time needed to solve the Vandermonde system. Evaluation time is similar to that of the Lagrange basis, but since there is no denominator numerical issues are less likely to appear.

We now have three strategies of interpolating k data points using a degree $k - 1$ polynomial by writing it in the monomial, Lagrange, and Newton bases. All three represent different compromises between numerical quality and speed, but the resulting interpolated function $f(x)$ is the *same* in each case. More explicitly, there is exactly one polynomial of degree $k - 1$ going through a set of k points, so since all our interpolants are degree $k - 1$ they must have the same output.

12.1.2 Alternative Bases

Although polynomial functions are particularly amenable to mathematical analysis, there is no fundamental reason why an interpolation basis cannot consist of different types of functions. For example, a crowning result of Fourier analysis implies that many functions are well-approximated by linear combinations of trigonometric functions $\cos(kx)$ and $\sin(kx)$

for $k \in \mathbb{N}$. A construction like the Vandermonde matrix still applies in this case, and the Fast Fourier Transform algorithm (which merits a larger discussion) solves the resulting linear system with remarkable efficiency.

A smaller extension of the development in §12.1.1 is to *rational* functions of the form:

$$f(x) \equiv \frac{p_0 + p_1 x + p_2 x^2 + \cdots + p_m x^m}{q_0 + q_1 x + q_2 x^2 + \cdots + q_n x^n}$$

If we are given k pairs (x_i, y_i), then we will need $m + n + 1 = k$ for this function to be well-defined. One degree of freedom must be fixed to account for the fact that the same rational function can be expressed multiple ways by simultaneously scaling the numerator and the denominator.

Rational functions can have asymptotes and other features not achievable using only polynomials, so they can be desirable interpolants for functions that change quickly or have poles. Once m and n are fixed, the coefficients p_i and q_i still can be found using linear techniques by multiplying both sides by the denominator:

$$y_i(q_0 + q_1 x_i + q_2 x_i^2 + \cdots + q_n x_i^n) = p_0 + p_1 x_i + p_2 x_i^2 + \cdots + p_m x_i^m$$

For interpolation, the unknowns in this expression are the p's and q's.

The flexibility of rational functions, however, can cause some issues. For instance, consider the following example:

Example 12.3 (Failure of rational interpolation, [111] §2.2)**.** Suppose we wish to find a rational function $f(x)$ interpolating the following data points: $(0, 1)$, $(1, 2)$, $(2, 2)$. If we choose $m = n = 1$, then the linear system for finding the unknown coefficients is:

$$\begin{aligned} q_0 &= p_0 \\ 2(q_0 + q_1) &= p_0 + p_1 \\ 2(q_0 + 2q_1) &= p_0 + 2p_1 \end{aligned}$$

One nontrivial solution to this system is:

$$\begin{aligned} p_0 &= 0 & q_0 &= 0 \\ p_1 &= 2 & q_1 &= 1 \end{aligned}$$

This implies the following form for $f(x)$:

$$f(x) = \frac{2x}{x}$$

This function has a degeneracy at $x = 0$, and canceling the x in the numerator and denominator does not yield $f(0) = 1$ as we might desire.

This example illustrates a larger phenomenon. The linear system for finding the p's and q's can run into issues when the resulting denominator $\sum_\ell p_\ell x^\ell$ has a root at any of the fixed x_i's. It can be shown that when this is the case, no rational function exists with the fixed choice of m and n interpolating the given values. A typical partial resolution in this case is presented in [111], which suggests incrementing m and n alternatively until a nontrivial solution exists. From a practical standpoint, however, the specialized nature of these methods indicates that alternative interpolation strategies may be preferable when the basic rational methods fail.

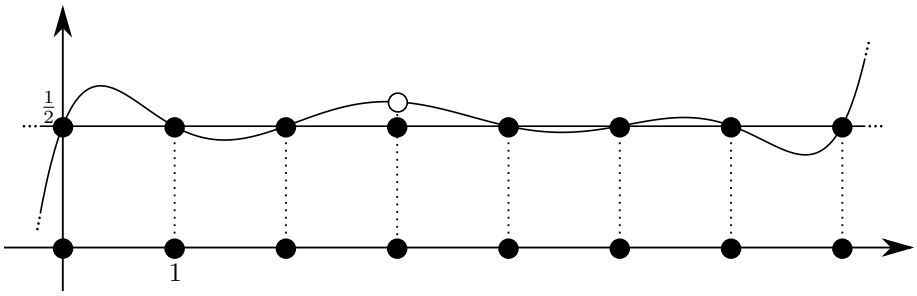


FIGURE 12.4 Interpolating eight samples of the function $f(x) \equiv 1/2$ using a seventh-degree polynomial yields a straight line, but perturbing a single data point at $x = 3$ creates an interpolant that oscillates far away from the infinitesimal vertical displacement.

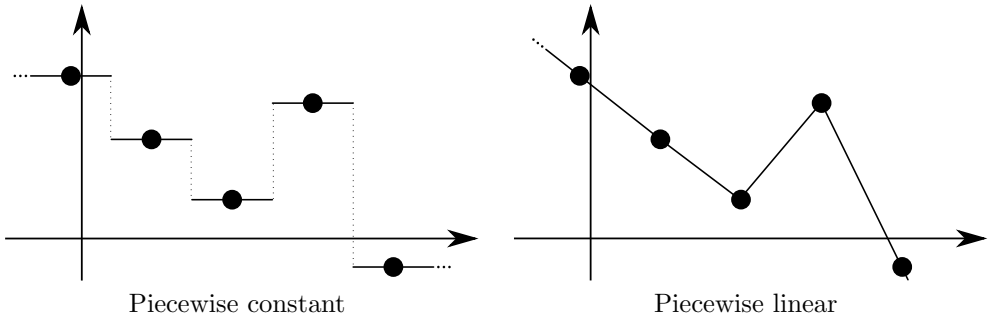


FIGURE 12.5 Two piecewise interpolation strategies.

12.1.3 Piecewise Interpolation

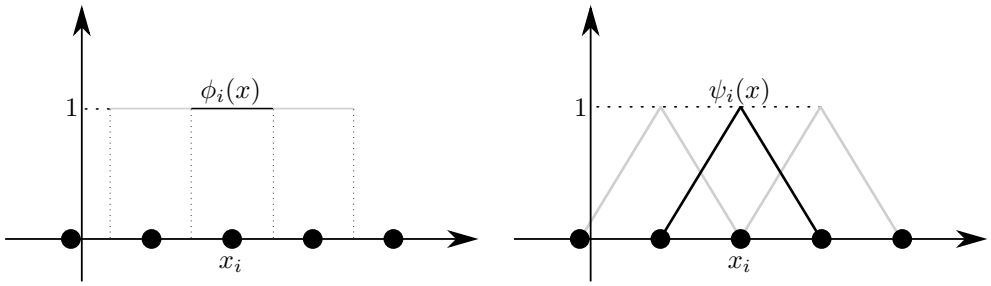
So far, we have constructed our interpolation strategies by combining simple functions defined on all of $\mathbb{R}$. When the number k of data points becomes high, however, many degeneracies become apparent. For example, Figure 12.4 illustrates how polynomial interpolation is *nonlocal*, meaning that changing any single value y_i in the input data can change the behavior of f for all x, even those that are far away from x_i. This property is undesirable for most applications: We usually expect only the input data near a given x to affect the value of the interpolated function $f(x)$, especially when there is a large cloud of input points.

For these reasons, when we design a set of basis functions $\phi_1, \ldots, \phi_k$, a desirable property we have not yet considered is that they have *compact support*:

Definition 12.1 (Compact support)**.** A function $g(\vec{x})$ has *compact support* if there exists $C \in \mathbb{R}$ such that $g(\vec{x}) = 0$ for any $\vec{x}$ with $\|\vec{x}\|_2 > C$.

That is, compactly-supported functions only have a finite range of points in which they can take nonzero values.

Piecewise formulas comprise the most common constructions for interpolatory bases with compact support. In particular, much of the literature on computer graphics depends on the construction of *piecewise polynomials*, which are defined by breaking $\mathbb{R}$ into a set of intervals and writing a different polynomial in each interval. To do so, we will order the



Piecewise constant basis Piecewise linear basis (hat function)

FIGURE 12.6 Basis functions corresponding to the piecewise interpolation strategies in Figure 12.5.

data points so that $x_1 < x_2 < \cdots < x_k$. Then, two examples of piecewise interpolants are the following, illustrated in Figure 12.5:

- Piecewise constant: For a given x, find the data point x_i minimizing $|x - x_i|$ and define $f(x) = y_i$.
- Piecewise linear: If $x < x_1$ take $f(x) = y_1$, and if $x > x_k$ take $f(x) = y_k$. Otherwise, find an interval with $x \in [x_i, x_{i+1}]$ and define

$$f(x) = y_{i+1} \cdot \frac{x - x_i}{x_{i+1} - x_i} + y_i \cdot \left(1 - \frac{x - x_i}{x_{i+1} - x_i}\right).$$

More generally, we can write a different polynomial in each interval $[x_i, x_{i+1}]$. Notice our pattern so far: Piecewise constant polynomials are discontinuous, while piecewise linear functions are continuous. Piecewise quadratics can be C^1, piecewise cubics can be C^2, and so on. This increased continuity and differentiability occurs even though each y_i has local support; this theory is worked out in detail in constructing "splines," or curves interpolating between points given function values and tangents.

This increased continuity, however, has its own drawbacks. With each additional degree of differentiability, we put a stronger smoothness assumption on f. This assumption can be unrealistic: Many physical phenomena truly are noisy or discontinuous, and this increased smoothness can negatively affect interpolatory results. One domain in which this effect is particularly clear is when interpolation is used in conjunction with physics simulation tools. Simulating turbulent fluid flows with oversmoothed functions can remove discontinuous phenomena like shock waves that are desirable as output.

These issues aside, piecewise polynomials still can be written as linear combinations of basis functions. For instance, the following functions serve as a basis for the piecewise constant functions:

$$\phi_i(x) = \begin{cases} 1 & \text{when } \frac{x_{i-1}+x_i}{2} \le x < \frac{x_i+x_{i+1}}{2} \\ 0 & \text{otherwise} \end{cases}$$

This basis puts the constant 1 near x_i and 0 elsewhere; the piecewise constant interpolation of a set of points (x_i, y_i) is written as $f(x) = \sum_i y_i \phi_i(x)$. Similarly, the so-called "hat" basis spans the set of piecewise linear functions with sharp edges at the data points x_i:

$$\psi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{when } x_{i-1} < x \le x_i \\ \frac{x_{i+1} - x}{x_{i+1} - x_i} & \text{when } x_i < x \le x_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

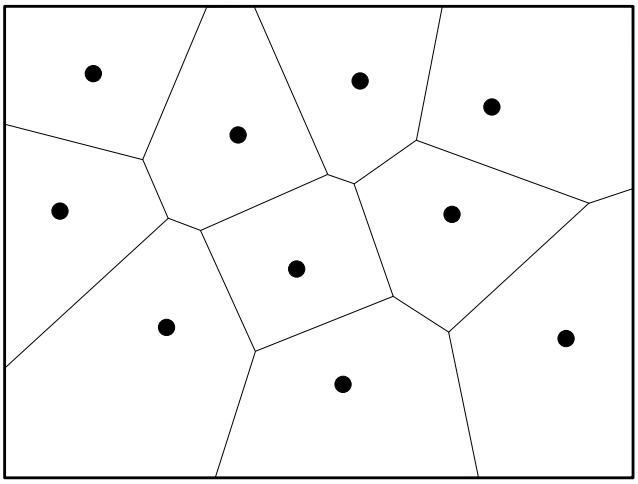


FIGURE 12.7 Voronoi cells associated with ten points in a rectangle.

Once again, by construction the piecewise linear interpolation of the given data points is $f(x) = \sum_i y_i \psi_i(x)$. Examples of both bases are shown in Figure 12.6.

12.2 MULTIVARIABLE INTERPOLATION

It is possible to extend the strategies above to the case of interpolating a function given data points $(\vec{x}_i, y_i)$ where $\vec{x}_i \in \mathbb{R}^n$ now can be multidimensional. Strategies for interpolation in this more general case are more challenging to formulate, however, because it is less obvious to partition $\mathbb{R}^n$ into a small number of regions around the source points $\vec{x}_i$.

12.2.1 Nearest-Neighbor Interpolation

Given the complication of interpolation on $\mathbb{R}^n$, a common pattern is to interpolate using many *low-order* functions rather than fewer smooth functions, that is, to prefer simplistic and efficient interpolation strategies over ones that output C^∞ functions. For example, if all we are given is a set of pairs $(\vec{x}_i, y_i)$, then one piecewise constant strategy for interpolation is to use *nearest-neighbor interpolation*. In this case, $f(\vec{x})$ takes the value y_i corresponding to $\vec{x}_i$ minimizing $\|\vec{x} - \vec{x}_i\|_2$. Simple implementations iterate over all i to find the closest $\vec{x}_i$ to $\vec{x}$, and data structures like k-d trees can find nearest neighbors more quickly.

Just as piecewise constant interpolation on $\mathbb{R}$ yielded constant values of $f(x)$ on intervals about the data points x_i, nearest-neighbor interpolation yields a function that is piecewise constant on *Voronoi cells*:

Definition 12.2 (Voronoi cell)**.** Given a set of points $S = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k\} \subseteq \mathbb{R}^n$, the *Voronoi cell* corresponding to a specific $\vec{x}_i \in S$ is the set $V_i \equiv \{\vec{x} : \|\vec{x} - \vec{x}_i\|_2 < \|\vec{x} - \vec{x}_j\|_2 \text{ for all } j \neq i\}$. That is, it is the set of points closer to $\vec{x}_i$ than to any other $\vec{x}_j$ in S.

Figure 12.7 shows an example of the Voronoi cells about a set of data points in $\mathbb{R}^2$. These cells have many favorable properties; for example, they are convex polygons and are localized about each $\vec{x}_i$. The adjacency of Voronoi cells is a well-studied problem in computational geometry leading to the construction of the celebrated Delaunay triangulation [30].

In many cases, however, it is desirable for the interpolant $f(\vec{x})$ to be continuous or

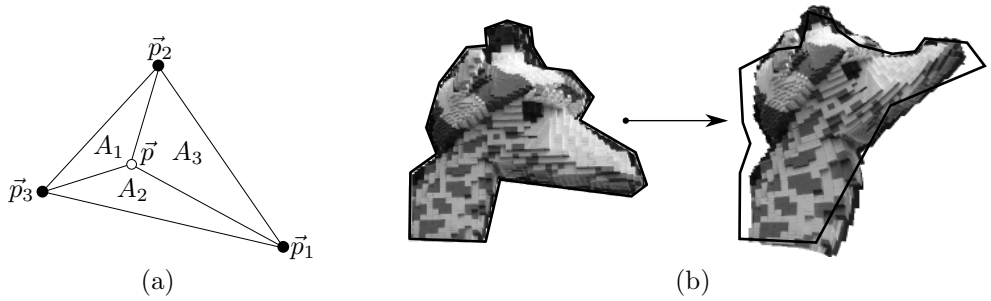


FIGURE 12.8 (a) The barycentric coordinates of $\vec{p} \in \mathbb{R}^2$ relative to the points $\vec{p}_1$, $\vec{p}_2$, and $\vec{p}_3$, resp., are $(A_1/A, A_2/A, A_3/A)$, where $A \equiv A_1 + A_2 + A_3$ and A_i is the area of triangle i; (b) the barycentric deformation method [123] uses a generalized version of barycentric coordinates to deform planar shapes according to motions of a polygon with more than three vertices.

differentiable. There are many options for continuous interpolation in $\mathbb{R}^n$, each with its own advantages and disadvantages. If we wish to extend the nearest-neighbor formula, we could compute multiple nearest neighbors $\vec{x}_1, \ldots, \vec{x}_k$ of $\vec{x}$ and interpolate $f(\vec{x})$ by averaging the corresponding $y_1, \ldots, y_k$ with distance-based weights; problem 12.3 explores one such weighting. Certain "k-nearest neighbor" data structures also can accelerate queries searching for multiple points in a dataset closest to a given $\vec{x}$.

12.2.2 Barycentric Interpolation

Another continuous multi-dimensional interpolant appearing frequently in the computer graphics literature is *barycentric* interpolation. Suppose we have exactly $n+1$ sample points $(\vec{x}_1, y_1), \ldots, (\vec{x}_{n+1}, y_{n+1})$, where $\vec{x}_i \in \mathbb{R}^n$, and we wish to interpolate the values y_i to all of $\mathbb{R}^n$; on the plane $\mathbb{R}^2$, we would be given three values associated with the vertices of a triangle. In the absence of degeneracies (e.g. three of the $\vec{x}_i$'s coinciding on the same line), any $\vec{x} \in \mathbb{R}^n$ can be written uniquely as a linear combination $\vec{x} = \sum_{i=1}^{n+1} a_i \vec{x}_i$ where $\sum_i a_i = 1$. That is, we write $\vec{x}$ as a weighted average of the $\vec{x}_i$'s with weights a_i. Barycentric interpolation then defines $f(\vec{x}) \equiv \sum_i a_i(\vec{x}) y_i$.

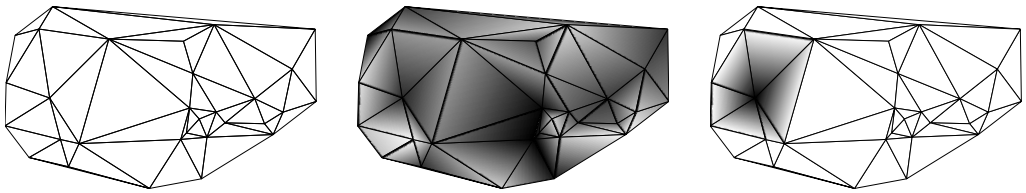
On the plane, barycentric interpolation has a straightforward geometric interpretation involving triangle areas, illustrated in Figure 12.8(a). Regardless of dimension, however, the barycentric interpolant $f(\vec{x})$ is *affine*, meaning it can be written $f(\vec{x}) = c + \vec{d} \cdot x$ for some $c \in \mathbb{R}$ and $\vec{d} \in \mathbb{R}^n$. Counting degrees of freedom, the $n+1$ sample points are accounted for via n unknowns in $\vec{d}$ and one unknown in c.

The system of equations solved for $\vec{a}$ to carry out barycentric interpolation at $\vec{x} \in \mathbb{R}^n$ is:

$$\sum_i a_i \vec{x}_i = \vec{x} \qquad \sum_i a_i = 1$$

As we discussed, this system usually is invertible when there are $n+1$ points $\vec{x}_i$. In the presence of additional $\vec{x}_i$'s, however, the system becomes *underdetermined*. This means that there are multiple ways of writing $\vec{x}$ as a weighted average of the $\vec{x}_i$'s, making room for additional design decisions for interpolation techniques.

One resolution of this non-uniqueness is to add more conditions on weights $\vec{a}$. These yield different *generalized barycentric coordinates*, a topic of research in mathematics and



(a) Triangle mesh (b) Barycentric interpolation (c) Hat function

FIGURE 12.9 (a) A collection of points on $\mathbb{R}^2$ can be triangulated into a triangle mesh; (b) using this mesh, a per-point function can be interpolated to the interior using per-triangle barycentric interpolation; (c) a single "hat" basis function takes value one on a single vertex and is interpolated using barycentric coordinates to the remainder of the domain.

engineering. Typical constraints on $\vec{a}$ ask that it is smooth as a function of $\vec{x}$ on $\mathbb{R}^n$ and nonnegative on the interior of the polygon or polyhedron bordered by the $\vec{x}_i$'s. Figure 12.8(b) shows an example of image deformation using a fairly recent generalized barycentric coordinates algorithm; the particular method shown makes use of complex-valued coordinates to take advantage of geometric properties of the complex plane [123].

Another way to carry out barycentric interpolation with more than $n+1$ data points employs *piecewise* affine functions for interpolation; we will restrict our discussion to $\vec{x}_i \in \mathbb{R}^2$ for simplicity, although extensions to higher dimensions are possible. Suppose we are given not only a set of points $\vec{x}_i \in \mathbb{R}^2$ but also a triangulation linking those points together, as in Figure 12.9(a); if the triangulation is not known *a priori* it can be computed using well-known geometric techniques [30]. Then, we can interpolate values from the vertices of each triangle to its interior using barycentric interpolation.

Example 12.4 (Shading)**.** A typical representation of three-dimensional shapes in computer graphics is a set of triangles linked into a mesh. In the *per-vertex* shading model, one color is computed for each vertex on the mesh using lighting of the scene, material properties, and so on. Then, to render the shape on-screen, those per-vertex colors are interpolated using barycentric interpolation to the interiors of the triangles. Similar strategies are used for texturing and other common tasks. Figure 12.9(b) shows an example of this technique.

As an aside, one pertinent issue specific to computer graphics is the interplay between perspective transformations and interpolation. Barycentric interpolation of color along a triangulated 3D surface and then projection of that color onto the image plane is not the same as projecting the triangles to the image plane and subsequently interpolating color along the projected two-dimensional triangles. Hence, algorithms in this domain must use *perspective-corrected* interpolation strategies to account for this discrepancy during the rendering process.

Interpolation using a triangulation parallels the use of a piecewise-linear hat basis for one-dimensional functions, introduced in §12.1.3. Now, we can think of $f(\vec{x})$ as a linear combination $\sum_i y_i \phi_i(\vec{x})$, where each $\phi_i(\vec{x})$ is the piecewise affine function obtained by putting a 1 on $\vec{x}_i$ and 0 everywhere else, as in Figure 12.9(c).

Given a set of points in $\mathbb{R}^2$, the problem of triangulation is far from trivial, and analogous constructions in higher dimensions can scale poorly. Thus, when $n > 3$, nearest-neighbor or regression methods usually are preferable to barycentric interpolation.

12.2.3 Grid-Based Interpolation

Rather than using triangles, an alternative decomposition of the domain of f occurs when the points $\vec{x}_i$ occur on a regular grid. The following examples illustrate situations when this is the case:

Example 12.5 (Image processing)**.** As mentioned in the introduction, a typical digital photograph is represented as an $m \times n$ grid of red, green, and blue color intensities. We can think of these values as living on the lattice $\mathbb{Z} \times \mathbb{Z}$. Suppose we wish to rotate the image by an angle that is not a multiple of 90°, however. Then, we must look up color values at potentially non-integer positions, requiring the interpolation of the image to $\mathbb{R} \times \mathbb{R}$.

Example 12.6 (Medical imaging)**.** The output of a magnetic resonance imaging (MRI) device is an $m \times n \times p$ grid of values representing the density of tissue at different points; a theoretical model for this data is as a function $f : \mathbb{R}^3 \to \mathbb{R}$. We can extract the outer surface of a particular organ by finding the level set $\{\vec{x} : f(\vec{x}) = c\}$ for some c. Finding this level set requires us to extend f to the entire voxel grid to find exactly where it crosses c.

Grid-based interpolation applies the one-dimensional formulae from §12.1.3 one dimension at a time. For example, *bilinear* interpolation in $\mathbb{R}^2$ applies linear interpolation in x_1 and then x_2 (or vice versa):

Example 12.7 (Bilinear interpolation)**.** Suppose f takes on the following values:

$$f(0,0) = 1 \qquad f(0,1) = -3 \qquad f(1,0) = 5 \qquad f(1,1) = -11$$

and that in between f is obtained by bilinear interpolation. To find $f(\frac{1}{4}, \frac{1}{2})$, we first interpolate in x_1 to find:

$$f\left(\frac{1}{4}, 0\right) = \frac{3}{4} f(0,0) + \frac{1}{4} f(1,0) = 2$$
$$f\left(\frac{1}{4}, 1\right) = \frac{3}{4} f(0,1) + \frac{1}{4} f(1,1) = -5$$

Next, we interpolate in x_2:

$$f\left(\frac{1}{4}, \frac{1}{2}\right) = \frac{1}{2} f\left(\frac{1}{4}, 0\right) + \frac{1}{2} f\left(\frac{1}{4}, 1\right) = -\frac{3}{2}$$

We receive the same output interpolating first in x_2 and second in x_1.

Higher-order methods like bicubic and Lanczos interpolation use more polynomial terms but are slower to evaluate. For example, bicubic interpolation requires values from more grid points than just the four function closest to $\vec{x}$ needed for bilinear interpolation. This additional expense can slow down image processing tools for which every lookup in memory incurs significant computation time.

12.3 THEORY OF INTERPOLATION

So far, our treatment of interpolation has been fairly heuristic. While relying on our intuition for what a "reasonable" interpolation for a set of function values for the most part

is acceptable, subtle issues can arise with different interpolation methods that should be acknowledged.

12.3.1 Linear Algebra of Functions

We began our discussion by posing interpolation strategies using different bases for the set of functions $f : \mathbb{R} \to \mathbb{R}$. This analogy of to vector spaces extends to a complete linear-algebraic theory of functions, and in many ways the field of *functional analysis* essentially extends the geometry of $\mathbb{R}^n$ to sets of functions. Here we will discuss functions of one variable, although many aspects of the extension to more general functions are easy to carry out.

Just as we can define notions of span and linear combination for functions, for fixed $a, b \in \mathbb{R}$ we also can define an *inner product* of functions $f(x)$ and $g(x)$ as follows:

$$\langle f, g \rangle \equiv \int_a^b f(x)g(x)\,dx.$$

We then can define a norm of a function to be $\|f\|_2 \equiv \sqrt{\langle f, f \rangle}$. These constructions are parallel to the corresponding constructions in $\mathbb{R}^n$; in particular, both the dot product $\vec{x}\cdot\vec{y}$ and the inner product $\langle f, g \rangle$ are obtained by multiplying the "elements" of the two multiplicands and summing—or integrating.

Example 12.8 (Function inner product)**.** Take $p_n(x) = x^n$ to be the n-th monomial. Then, for $a = 0$ and $b = 1$ we have:

$$\begin{aligned}\langle p_n, p_m \rangle &= \int_0^1 x^n \cdot x^m\,dx \\ &= \int_0^1 x^{n+m}\,dx \\ &= \frac{1}{n+m+1}\end{aligned}$$

This shows:

$$\begin{aligned}\left\langle \frac{p_n}{\|p_n\|}, \frac{p_m}{\|p_m\|} \right\rangle &= \frac{\langle p_n, p_m \rangle}{\|p_n\|\|p_m\|} \\ &= \frac{\sqrt{(2n+1)(2m+1)}}{n+m+1}\end{aligned}$$

This value is approximately 1 when $n \approx m$ but $n \neq m$, substantiating our earlier claim illustrated in Figure 12.1 that the monomials "overlap" considerably on $[0, 1]$.

Given this inner product, we can apply the Gram-Schmidt algorithm to find an orthonormal basis for the set of polynomials. If we take $a = -1$ and $b = 1$, applying Gram-Schmidt to the monomial basis yields the Legendre polynomials, plotted in Figure 12.10:

$$\begin{aligned}P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x)\end{aligned}$$

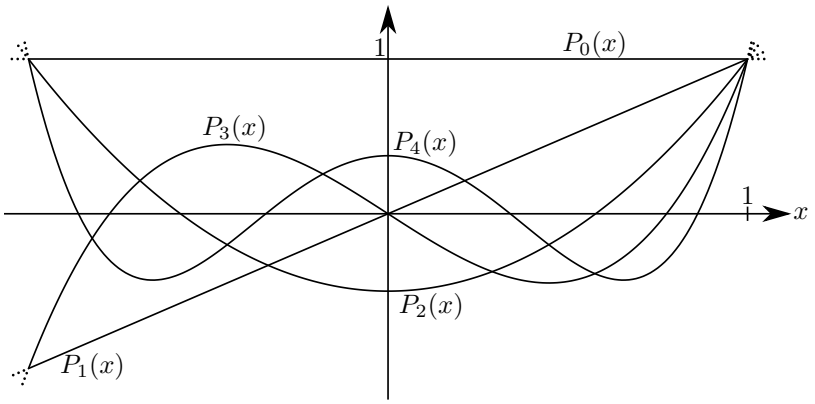


FIGURE 12.10 The first five Legendre polynomials, notated $P_0(x), \ldots, P_4(x)$.

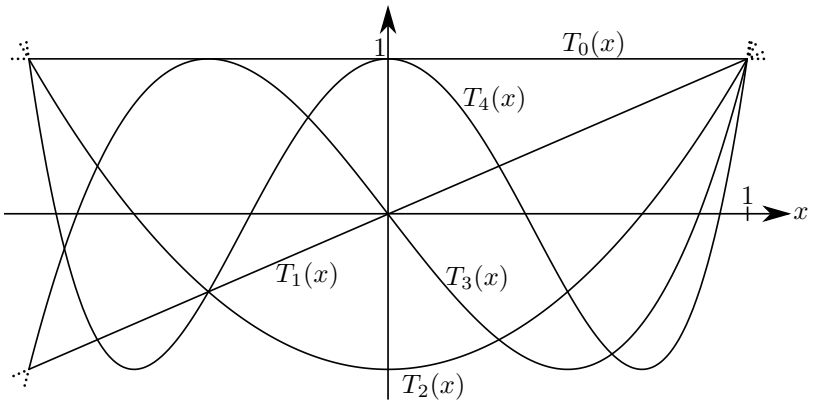


FIGURE 12.11 The first five Chebyshev polynomials, notated $T_0(x), \ldots, T_4(x)$.

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$
$$\vdots \qquad \vdots$$

These polynomials have many useful properties thanks to their orthogonality. For example, suppose we wish to approximate $f(x)$ with a sum $\sum_i a_i P_i(x)$. If we wish to minimize $\|f - \sum_i a_i P_i\|_2$ in the functional norm, this is a *least squares* problem! By orthogonality of the Legendre basis for $\mathbb{R}[x]$, our formula from Chapter 4 for projection onto an orthogonal basis shows:

$$a_i = \frac{\langle f, P_i \rangle}{\langle P_i, P_i \rangle}$$

Thus, approximating f using polynomials can be accomplished by integrating f against the members of the Legendre basis; in the next chapter we will learn how this integral can be carried out numerically.

Given a positive function $w(x)$, we can define a more general inner product $\langle \cdot, \cdot \rangle_w$ by writing

$$\langle f, g \rangle_w \equiv \int_a^b w(x) f(x) g(x)\, dx.$$

If we take $w(x) = \frac{1}{\sqrt{1-x^2}}$ with $a = -1$ and $b = 1$, then Gram-Schmidt on the monomials yields the *Chebyshev* polynomials, shown in Figure 12.11:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ &\vdots \quad \vdots \end{aligned}$$

A surprising identity holds for these polynomials:

$$T_k(x) = \cos(k \arccos(x)).$$

This formula can be checked by explicitly checking it for T_0 and T_1, and then inductively applying the observation:

$$\begin{aligned} T_{k+1}(x) &= \cos((k+1)\arccos(x)) \\ &= 2x\cos(k\arccos(x)) - \cos((k-1)\arccos(x)) \text{ by the identity} \\ &\qquad\qquad \cos((k+1)\theta) = 2\cos(k\theta)\cos(\theta) - \cos((k-1)\theta) \\ &= 2xT_k(x) - T_{k-1}(x) \end{aligned}$$

This "three-term recurrence" formula also gives an easy way to generate expressions for the Chebyshev polynomials.

Thanks to this trigonometric characterization of the Chebyshev polynomials, the minima and maxima of T_k oscillate between $+1$ and -1. Furthermore, these extrema are located at $x = \cos(i\pi/k)$ (the so-called "Chebyshev points") for i from 0 to k; this nice distribution of extrema avoids oscillatory phenomena like that shown in Figure 12.4 when using a finite number of polynomial terms to approximate a function. More technical treatments of polynomial interpolation recommend placing samples x_i for interpolation near Chebyshev points to obtain smooth output.

12.3.2 Approximation via Piecewise Polynomials

Suppose we wish to approximate a function $f(x)$ with a polynomial of degree n on an interval $[a, b]$. Define Δx to be the spacing $b - a$. One measure of error of an approximation is as a function of Δx, which should vanish as $\Delta x \to 0$. Then, if we approximate f with piecewise polynomials, this type of analysis tells us how far apart we should space the polynomials to achieve a desired level of approximation.

For example, suppose we approximate f with a constant $c = f(\frac{a+b}{2})$, as in piecewise constant interpolation. If we assume $|f'(x)| < M$ for all $x \in [a, b]$, we have:

$$\begin{aligned} \max_{x\in[a,b]} |f(x) - c| &\leq \Delta x \max_{x\in[a,b]} M \text{ by the mean value theorem} \\ &\leq M\Delta x \end{aligned}$$

Thus, we expect $O(\Delta x)$ error when using piecewise constant interpolation.

Suppose instead we approximate f using piecewise linear interpolation, that is, by taking

$$\tilde{f}(x) = \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b).$$

By the mean value theorem, we know $\tilde{f}'(x) = f'(\theta)$ for some $\theta \in [a, b]$. Writing the Taylor expansion about θ shows $f(x) = f(\theta) + f'(\theta)(x - \theta) + O(\Delta x^2)$ on $[a, b]$, while we can rewrite our linear approximation as $\tilde{f}(x) = f(\theta) + f'(\theta)(x - \theta)$. Thus, subtracting these two expressions shows that the approximation error of f decreases to $O(\Delta x^2)$. It is not difficult to predict that approximation with a degree n polynomial makes $O(\Delta x^{n+1})$ error, although in practice the quadratic convergence of piecewise linear approximations suffices for most applications.

12.4 EXERCISES

12.1 Write the degree-four polynomial interpolating between the data points $(-2, 15)$, $(0, -1)$, $(1, 0)$, and $(3, -2)$.
Hint: Your answer does not have to be written in the monomial basis.

12.2 Show that the interpolation from Example 12.7 yields the same result regardless of whether x_1 or x_2 is interpolated first.

12.3 ("Inverse distance weighting") Suppose we are given a set of distinct points $\vec{x}_1, \ldots, \vec{x}_k \in \mathbb{R}^n$ with labels $y_1, \ldots, y_k \in \mathbb{R}$. Then, one interpolation strategy defines an interpolant $f(\vec{x})$ as follows [103]:

$$f(\vec{x}) \equiv \begin{cases} y_i & \text{if } \vec{x} = \vec{x}_i \text{ for some } i \\ \frac{\sum_i w_i(\vec{x}) y_i}{\sum_i w_i(\vec{x})} & \text{otherwise} \end{cases},$$

where $w_i(\vec{x}) \equiv \|\vec{x} - \vec{x}_i\|_2^{-p}$ for some fixed $p \geq 1$.

(a) Argue that as $p \to \infty$, the interpolant $f(\vec{x})$ becomes piecewise constant on the Voronoi cells of the $\vec{x}_i$'s.

(b) Define the function

$$\phi(\vec{x}, y) \equiv \left(\sum_i \frac{(y - y_i)^2}{\|\vec{x} - \vec{x}_i\|_2^p} \right)^{1/p}.$$

Show that for fixed $\vec{x} \in \mathbb{R}^n$, the value $f(\vec{x})$ is the minimum of $\phi(\vec{x}, y)$ over all y.

(c) Evaluating the sum in this formula can be expensive when k is large. Propose a modification to the w_i's that avoids this issue; there are many possible techniques here.

12.4 ("Cubic Hermite interpolation") In computer graphics, a common approach to drawing curves is to use cubic interpolation. Typically, artists design curves by specifying their endpoints as well as the tangents to the curves at the endpoints.

(a) Suppose $P(t)$ is the cubic polynomial:

$$P(t) = at^3 + bt^2 + ct + d.$$

Write a set of linear conditions on a, b, c, and d such that $P(t)$ satisfies the following conditions for fixed values of h_0, h_1, h_2, and h_3:

$$\begin{aligned} P(0) &= h_0 & \qquad P'(0) &= h_2 \\ P(1) &= h_1 & \qquad P'(1) &= h_3. \end{aligned}$$

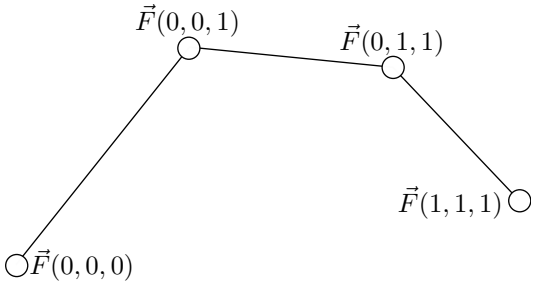


FIGURE 12.12 Diagram for problem 12.5d.

(b) Write the *cubic Hermite* basis for cubic polynomials $\{\phi_0(t), \phi_1(t), \phi_2(t), \phi_3(t)\}$ such that $P(t)$ satisfying the conditions from 12.4a can be written

$$P(t) = h_0\phi_0(t) + h_1\phi_1(t) + h_2\phi_2(t) + h_3\phi_3(t).$$

12.5 ("Cubic blossom") We continue to explore interpolation techniques suggested in the previous problem.

(a) Given $P(t) = at^3 + bt^2 + ct + d$, define a *cubic blossom function* $F(t_1, t_2, t_3)$ in terms of $\{a, b, c, d\}$ satisfying the following properties [97]:

Symmetric: $F(t_1, t_2, t_3) = F(t_i, t_j, t_k)$ for any permutation (i, j, k) of $\{1, 2, 3\}$

Affine: $F(\alpha u + (1-\alpha)v, t_2, t_3) = \alpha F(u, t_2, t_3) + (1-\alpha)F(v, t_2, t_3)$

Diagonal: $f(t) = F(t, t, t)$

(b) Now, define

$$p = F(0,0,0) \qquad q = F(0,0,1)$$
$$r = F(0,1,1) \qquad s = F(1,1,1).$$

Write expressions for $f(0)$, $f(1)$, $f'(0)$, and $f'(1)$ in terms of p, q, r, and s.

(c) Write a basis $\{B_0(t), B_1(t), B_2(t), B_3(t)\}$ for cubic polynomials such that given a cubic blossom $F(t_1, t_2, t_3)$ of $f(t)$ we can write

$$f(t) = F(0,0,0)B_0(t) + F(0,0,1)B_1(t) + F(0,1,1)B_2(t) + F(1,1,1)B_3(t).$$

The functions $B_i(t)$ are known as the cubic Bernstein basis.

(d) Suppose $F_1(t_1, t_2, t_3)$ and $F_2(t_1, t_2, t_3)$ are the cubic blossoms of functions $f_1(t)$ and $f_2(t)$, respectively, and define $\vec{F}(t_1, t_2, t_3) \equiv (F_1(t_1, t_2, t_3), F_2(t_1, t_2, t_3))$. Consider the four points shown in Figure 12.12. By bisecting line segments and drawing new ones, show how to construct $\vec{F}(1/2, 1/2, 1/2)$.

12.6[DH] Consider the polynomial $p(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n$. Alternatively, we can write $p(x)$ in the Newton basis relative to $x_0, \ldots, x_{n-1}$ as

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n \prod_{i=0}^{n-1} (x - x_i),$$

where $x_0, \ldots, x_{n-1}$ are fixed constants.

(a) Argue why we can write any n-th degree $p(x)$ in this form.

(b) Find explicit expressions for a_0, a_1, and a_2 in terms of x_0, x_1, and evaluations of $p(\cdot)$. Based on these expressions (and computing more terms if needed), propose a pattern for finding a_k.

(c) Use function evaluation to define the *zeroth divided difference* of p as $p[x_0] = p(x_0)$. Furthermore, define the *first divided difference* of p as

$$p[x_0, x_1] = \frac{p[x_0] - p[x_1]}{x_0 - x_1}.$$

Finally, define the *second divided difference* as

$$p[x_0, x_1, x_2] = \frac{p[x_0, x_1] - p[x_1, x_2]}{x_0 - x_2}.$$

Based on this pattern and the pattern you observed in the previous part, define the k-th divided difference of p.

(d) Write $p(x)$ in terms of the Newton basis and the divided differences.

(e) Suppose add another point (x_n, y_n) and wish to recompute the Newton interpolant. How many Newton coefficients need to be recomputed? Why?

12.7 ("Horner's rule") Consider the polynomial $p(x) \equiv a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k$. For fixed $x_0 \in \mathbb{R}$, define $c_0, \ldots, c_k \in \mathbb{R}$ recursively as follows:

$$c_k \equiv a_k$$
$$c_i \equiv a_i + c_{i+1} x_0 \ \forall i < k$$

Show $c_0 = p(x_0)$, and compare the number of multiplication and addition operations needed to compute $p(x_0)$ using this method versus the formula in terms of the a_i's.

12.8[DH] Consider the L_2 distance between polynomials f, g on $[-1, 1]$, given by

$$||f - g||_2 \equiv \left[\int_{-1}^{1} |f(x) - g(x)|^2 \, dx\right]^{1/2},$$

which arises from the inner product $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,dx$. Let $\mathcal{P}_n$ be the vector space of polynomials of degree no more than n, endowed with the above inner product. As we have discussed, polynomials $\{p_i\}_{i=1}^{m}$ are *orthogonal* with respect to this inner product if for all $i \neq j$, $\langle p_i, p_j\rangle = 0$; we can systematically obtain a set of orthogonal polynomials using the Gram-Schmidt process.

(a) Derive the first four Legendre polynomials via Gram-Schmidt orthogonalization of the monomials $1, x, x^2, x^3$.

(b) Suppose we wish to approximate a function f with a polynomial g. To do so, we can find the $g \in \mathcal{P}_n$ that is the best least-squares fit for f. Given the above discussion, write an optimization problem for finding g.

(c) Suppose we construct the *Gram matrix* G with entries $g_{ij} \equiv \langle p_i, p_j\rangle$ for a basis of polynomials $p_1, \ldots, p_n \in \mathcal{P}_n$. How is G involved in solving part 12.8b? What is the structure of G when $p_1, \ldots, p_n$ are the first n Legendre polynomials?

12.9[DH] For a given n, the *Chebyshev points* are given by $x_k = \cos\left(\frac{k\pi}{n}\right)$, where $k \in \{0, \ldots, n\}$.

(a) Show that the Chebyshev points are the projections onto the real line of n *evenly-spaced* points on the upper half of the unit circle in the complex plane.
Hint: Use complex exponentials.

(b) Suppose rather than proving the identity we *define* the Chebyshev polynomials using the expression $T_k(x) \equiv \cos(k \arccos(x))$. Starting from this expression, compute the first four Chebyshev polynomials in the monomial basis.

(c) Show that the Chebyshev polynomials you computed in the previous part are orthogonal with respect to the inner product $\langle f, g \rangle \equiv \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}} dx$.

(d) Compute the Chebyshev points for $n = 1, 2, 3$ and show that they are the local extrema of $T_1(x), T_2(x)$, and $T_2(x)$.

12.10 We can use interpolation strategies to formulate methods for root-finding in one or more variables.

(a) Find expressions for parameters a, b, c of the *linear fractional transformation*

$$f(x) \equiv \frac{x + a}{bx + c}$$

going through the points (x_0, y_0), (x_1, y_1), and (x_2, y_2).

(b) Find x_4 such that $f(x_4) = 0$; write x_4 in terms of the values (x_i, y_i) from the previous part.

(c) Suppose we are given a function $f(x)$ and wish to find a root x^* with $f(x^*) = 0$. Suggest an algorithm for root-finding using the construction in part 12.10b.

CHAPTER 13

Integration and Differentiation

CONTENTS

13.1 Motivation 270
13.2 Quadrature 271
13.2.1 Interpolatory Quadrature 272
13.2.2 Quadrature Rules 273
13.2.3 Newton-Cotes Quadrature 274
13.2.4 Gaussian Quadrature 278
13.2.5 Adaptive Quadrature 279
13.2.6 Multiple Variables 281
13.2.7 Conditioning 282
13.3 Differentiation 282
13.3.1 Differentiating Basis Functions 283
13.3.2 Finite Differences 283
13.3.3 Richardson Extrapolation 285
13.3.4 Choosing the Step Size 286
13.3.5 Automatic Differentiation 287
13.3.6 Integrated Quantities and Structure Preservation 288

IN the previous chapter, we developed tools for filling in reasonable values of a function $f(\vec{x})$ given a sampling of points $(\vec{x}_i, f(\vec{x}_i))$ in the domain of f. Such interpolation methods are useful in themselves for completing functions that are known to be continuous or differentiable but whose values only are sampled at a set of isolated points, but in some cases we then wish to study properties of these functions. In particular, if we wish to apply tools from calculus to f, we must be able to approximate its integrals and derivatives.

There are many applications in which numerical integration and differentiation play key roles for computation. In the most straightforward instance, some well-known functions are *defined* as integrals. For instance, the "error function" given by the cumulative distribution of a bell curve is defined as:

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

Approximations of $\operatorname{erf}(x)$ are needed in statistical methods, and one reasonable approach to finding these values is to compute the integral above numerically.

Other times, numerical approximations of derivatives and integrals are part of a larger system. For example, methods we will develop in future chapters for approximating solutions

to differential equations will depend strongly on discretizations of derivatives. In computational electrodynamics, *integral equations* that must be solved for an unknown function $\phi(\vec{y})$ given a kernel $K(\vec{x}, \vec{y})$ and function $f(\vec{x})$ appear in the relationship:

$$f(\vec{x}) = \int_{\mathbb{R}^n} K(\vec{x}, \vec{y})\phi(\vec{y})\, d\vec{y}.$$

These equations are solved to estimate electric and magnetic fields, but unless the ϕ and K are very special we cannot hope to work with such an integral in closed form, let alone solve this equation for the function $\phi(\vec{y})$.

In this chapter, we will develop methods for numerical integration and differentiation given a sampling of function values. We also will develop strategies to evaluate how well we expect approximations of derivatives and integrals to perform, helping formalize our intuition for their relative quality and efficiency in different circumstances or applications.

13.1 MOTIVATION

It is not hard to encounter applications of numerical integration and differentiation, given how often the tools of calculus appear in physics, statistics, and other fields. Well-known formulas aside, here we suggest a few less obvious places requiring tools for integration and differentiation.

Example 13.1 (Sampling from a distribution)**.** Suppose we are given a probability distribution $p(t)$ on the interval $[0, 1]$; that is, if we randomly sample values according to this distribution, we expect $p(t)$ to be proportional to the number of times we draw a value near t. A common task is to generate random numbers distributed like $p(t)$.

Rather than develop a specialized sampling method every time we receive a new $p(t)$, it is possible to leverage a single uniform sampling tool to sample from nearly any distribution on $[0, 1]$.We define the *cumulative distribution function* (CDF) of p to be

$$F(t) = \int_0^t p(x)\, dx.$$

Then, if X is a random number distributed evenly in $[0, 1]$, one can show that $F^{-1}(X)$ is distributed like p, where F^{-1} is the inverse of F. That is, if we can approximate F or F^{-1}, we can generate random numbers according to an arbitrary distribution p.

Example 13.2 (Optimization)**.** Most of our methods for minimizing and finding roots of a function $f(\vec{x})$ require computing not only values $f(\vec{x})$ but also gradients $\nabla f(\vec{x})$ and even Hessians $H_f(\vec{x})$. BFGS and Broyden's method can build up rough approximations of the highest-order derivatives of f during optimization. When f contains high frequencies, however, it may be better to approximate ∇f directly near the current iterate $\vec{x}_k$ rather than using values from potentially far-away iterates $\vec{x}_\ell$ for $\ell < k$, which can happen as BFGS or Broyden slowly build up derivative matrices.

Example 13.3 (Rendering)**.** The *rendering equation* from computer graphics and ray tracing is an integral equation expressing conservation of light energy [65]. As it was originally presented, the rendering equation states:

$$I(\vec{x}, \vec{y}) = g(\vec{x}, \vec{y}) \left[\varepsilon(\vec{x}, \vec{y}) + \int_S \rho(\vec{x}, \vec{y}, \vec{z}) I(\vec{y}, \vec{z})\, d\vec{z} \right]$$

Here $I(\vec{x}, \vec{y})$ is proportional to the intensity of light going from point $\vec{y}$ to point $\vec{x}$ in a scene. The functions on the right hand side are:

$g(\vec{x}, \vec{y})$	A *geometry term* accounting e.g. for objects occluding the path from $\vec{x}$ to $\vec{y}$
$\varepsilon(\vec{x}, \vec{y})$	The *light emitted* directly from $\vec{x}$ to $\vec{y}$
$\rho(\vec{x}, \vec{y}, \vec{z})$	A *scattering* term giving the amount of light scattered to point $\vec{x}$ by a patch of surface at location $\vec{z}$ from light located at $\vec{z}$
$S = \cup_i S_i$	The set of surfaces S_i in the scene

Many rendering algorithms can be described as approximate strategies for solving this integral equation.

Example 13.4 (Image processing)**.** Suppose we think of an image or photograph as a function of two variables $I(x, y)$ giving the brightness of the image at each position (x, y). Many classical image processing filters, including Gaussian blurs, can be thought of as *convolutions*, given by

$$(I * g)(x, y) = \iint_{\mathbb{R}^2} I(u, v) g(x - u, y - v)\, du\, dv.$$

For example, to blur an image we can take g to be a Gaussian; in this case $(I * g)(x, y)$ is a weighted average of the colors of I near the point (x, y). In practice, however, images are sampled at discrete grids of pixels, so this integral must be approximated.

Example 13.5 (Bayes' Rule)**.** Suppose X and Y are continuously-valued random variables; we can use $P(X)$ and $P(Y)$ to express the probabilities that X and Y take particular values. Sometimes, knowing X may affect our knowledge of Y. For instance, if X is a patient's blood pressure and Y is a patient's weight, then knowing a patient has high weight may suggest that he or she also has high blood pressure. In this situation, we can write *conditional* probability distributions $P(X|Y)$ (read "the probability of X given Y") expressing such relationships.

A foundation of modern probability theory states that $P(X|Y)$ and $P(Y|X)$ are related as follows:

$$P(X|Y) = \frac{P(Y|X)P(X)}{\int P(Y|X)P(X)\, dY}$$

Estimating the integral in the denominator can be a serious problem in machine learning algorithms where the probability distributions take complex forms. Approximate and often randomized integration schemes are needed for algorithms in parameter selection that use this value as part of a larger optimization technique [58].

13.2 QUADRATURE

We will begin by considering the problem of numerical integration, or *quadrature*. This problem—in a single variable—can be expressed as, "Given a sampling of n points from some function $f(x)$, find an approximation of $\int_a^b f(x)\, dx$." In the previous section, we presented several situations that boil down to exactly this problem.

There are a few variations of this setup that require slightly different treatment or adaptation:

- The endpoints a and b may be fixed, or we may wish to find a quadrature scheme that efficiently can approximate integrals for many (a, b) pairs.

- We may be able to query $f(x)$ at any x but wish to approximate the integral using relatively few samples, or we may be given a list of precomputed pairs $(x_i, f(x_i))$ and are constrained to using these data points in our approximation.

These considerations should be kept in mind as we design assorted quadrature techniques.

13.2.1 Interpolatory Quadrature

Many of the interpolation strategies developed in the previous chapter can be extended to methods for quadrature. Suppose we write a function $f(x)$ in terms of a set of basis functions $\phi_i(x)$:

$$f(x) = \sum_i a_i \phi_i(x).$$

Then, we can find the integral of f as follows:

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_a^b \left[\sum_i a_i \phi_i(x)\right] dx \text{ by definition of } f \\
&= \sum_i a_i \left[\int_a^b \phi_i(x)\,dx\right] \text{ by swapping the sum and the integral} \\
&= \sum_i c_i a_i \text{ if we make the definition } c_i \equiv \int_a^b \phi_i(x)\,dx
\end{aligned}$$

In other words, the integral of $f(x)$ written in a basis is a weighted sum of the integrals of the basis functions making up f.

Example 13.6 (Monomials)**.** Suppose we write $f(x) = \sum_k a_k x^k$. We know

$$\int_0^1 x^k\,dx = \frac{1}{k+1}.$$

Applying the formula above, we can write

$$\int_0^1 f(x)\,dx = \sum_k \frac{a_k}{k+1}.$$

In the more general notation above, we have taken $c_k = \frac{1}{k+1}$. This formula shows that the integral of $f(x)$ in the monomial basis can be computed directly via a weighted sum of the coefficients a_k.

Integration schemes derived using interpolatory basis functions are known as *interpolatory quadrature* rules; nearly all the methods we will present below can be written this way. We can encounter a chicken-and-egg problem if the integral $\int \phi_i(x)\,dx$ itself is not known in closed form. Certain methods in higher-order finite elements deal with this problem by putting extra computational time into making a high-quality numerical approximation of the integral of a single ϕ_i. Then, since all the ϕ's have similar form, these methods apply change-of-coordinates formulas to write integrals for the remaining basis functions. The canonical integral can be approximated offline using a high-accuracy scheme and then reused during computations where timing matters.

13.2.2 Quadrature Rules

Our discussion above suggests the following form for a *quadrature rule* approximating the integral of f on some interval given a set of sample locations x_i:

$$Q[f] \equiv \sum_i w_i f(x_i).$$

Different weights w_i yield different approximations of the integral, which we hope become increasingly similar as the x_i's are sampled more densely. From this perspective, the choices of $\{x_i\}$ and $\{w_i\}$ determine a quadrature rule.

Even the classical theory of integration suggests that this formula is a reasonable starting point. For example, the *Riemann integral* presented in many introductory calculus texts takes the form:

$$\int_a^b f(x)\,dx \equiv \lim_{\Delta x_k \to 0} \sum_k f(\tilde{x}_k)(x_{k+1} - x_k)$$

Here, the interval $[a, b]$ is partitioned into pieces $a = x_1 < x_2 < \cdots < x_n = b$, where $\Delta x_k = x_{k+1} - x_k$, and $\tilde{x}_k$ is any point in $[x_k, x_{k+1}]$. For a fixed set of x_k's before taking the limit, this integral is in the $Q[f]$ form above.

There are many ways to choose the form of $Q[\cdot]$, as we will see in the coming section and as we already have seen for interpolatory quadrature. If we can query f for its values anywhere, then the x_i's *and* w_i's can be chosen strategically to sample f in a near-optimal way, but even if the x_i's are fixed there exist many ways to choose the weights w_i with different advantages and disadvantages.

Example 13.7 (Method of undetermined coefficients)**.** Suppose we fix $x_1, \ldots, x_n$ and wish to find a reasonable set of weights w_i so that $\sum_i w_i f(x_i)$ approximates the integral of f for reasonably smooth $f : [a, b] \to \mathbb{R}$. An alternative to interpolatory quadrature is the *method of undetermined coefficients*. In this strategy, we choose n functions $f_1(x), \ldots, f_n(x)$ whose integrals are known, and require that the quadrature rule recovers the integrals of these functions exactly:

$$\begin{aligned}
\int_a^b f_1(x)\,dx &= w_1 f_1(x_1) + w_2 f_1(x_2) + \cdots + w_n f_1(x_n) \\
\int_a^b f_2(x)\,dx &= w_1 f_2(x_1) + w_2 f_2(x_2) + \cdots + w_n f_2(x_n) \\
&\vdots \qquad\qquad \vdots \\
\int_a^b f_n(x)\,dx &= w_1 f_n(x_1) + w_2 f_n(x_2) + \cdots + w_n f_n(x_n)
\end{aligned}$$

The n expressions above create an $n \times n$ linear system of equations for the unknown w_i's.

One common choice is to take $f_k(x) \equiv x^{k-1}$, that is, to make sure that the quadrature scheme recovers the integrals of low-order polynomials. As in Example 13.6, we know

$$\int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{k+1}.$$

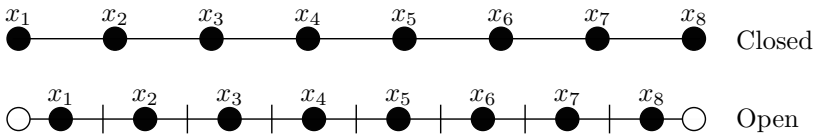


FIGURE 13.1 Closed and open Newton-Cotes quadrature schemes differ by where they place the samples x_i on the interval $[a, b]$; here we show the two samplings for $n = 8$.

Thus, we must solve the following linear system of equations for the w_i's:

$$\begin{aligned} w_1 + w_2 + \cdots + w_n &= b - a \\ x_1 w_1 + x_2 w_2 + \cdots + x_n w_n &= \frac{b^2 - a^2}{2} \\ x_1^2 w_1 + x_2^2 w_2 + \cdots + x_n^2 w_n &= \frac{b^3 - a^3}{3} \\ \vdots \qquad & \quad \vdots \\ x_1^{n-1} w_1 + x_2^{n-1} w_2 + \cdots + x_n^{n-1} w_n &= \frac{b^n - a^n}{n} \end{aligned}$$

In matrix form, this system is

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} b - a \\ \frac{1}{2}(b^2 - a^2) \\ \frac{1}{3}(b^3 - a^3) \\ \vdots \\ \frac{1}{n}(b^n - a^n) \end{pmatrix}$$

This is the transpose of the Vandermonde system discussed in §12.1.1.

13.2.3 Newton-Cotes Quadrature

Quadrature rules that integrate the result of polynomial interpolation when the $x_i's$ are evenly spaced in $[a, b]$ are known as *Newton-Cotes* quadrature rules. As illustrated in Figure 13.1, there are two reasonable choices of evenly-spaced samples:

- *Closed* Newton-Cotes quadrature places x_i's at a and b. In particular, for $k \in \{1, \ldots, n\}$ we take
$$x_k \equiv a + \frac{(k-1)(b-a)}{n-1}.$$
- *Open* Newton-Cotes quadrature does not place an x_i at a or b:
$$x_k \equiv a + \frac{k(b-a)}{n+1}.$$

The Newton-Cotes formulae compute the integral of the polynomial interpolant approximating the function on a to b through these points; the degree of the polynomial must be

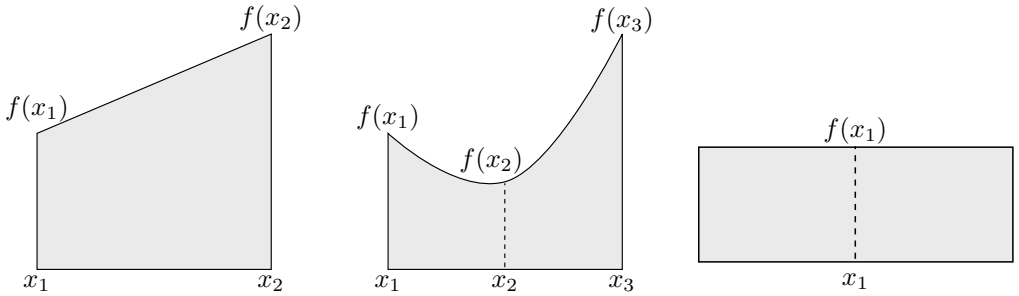


FIGURE 13.2 Newton-Cotes quadrature schemes; the approximated integral based on the $(x_i, f(x_i))$ pairs shown is given by the area of the gray region.

$n - 1$ to keep the quadrature rule well-defined. There is no inherent advantage to using closed versus open Newton-Cotes rules; the choice between these options generally depends on which set of samples is available.

We illustrate the integration rules below in Figure 13.2. We will keep n relatively small to avoid oscillation and noise sensitivity that occur when fitting high-degree polynomials to a set of data points. Then, as in piecewise polynomial interpolation, we will then chain together small pieces into *composite* rules when integrating over a large interval $[a, b]$.

Closed rules. Closed Newton-Cotes quadrature strategies require $n \geq 2$ to avoid dividing by zero. The two lowest-order closed integrators are the most common:

- The *trapezoidal* rule is obtained for $n = 2$ (so $x_1 = a$ and $x_2 = b$) by linearly interpolating from $f(a)$ to $f(b)$. It effectively computes the area of a trapezoid via the formula:
$$\int_a^b f(x)\,dx \approx (b-a)\frac{f(a)+f(b)}{2}.$$
- *Simpson's rule* comes from taking $n = 3$, with sample points
$$\begin{aligned} x_1 &= a \\ x_2 &= \frac{a+b}{2} \\ x_3 &= b. \end{aligned}$$
Integrating the parabola that goes through these three points yields
$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Open rules. By far the most common rule for open quadrature is the *midpoint rule*, which takes $n = 1$ and approximates an integral with the signed area of a rectangle through the midpoint of the integration interval $[a, b]$:
$$\int_a^b f(x)\,dx \approx (b-a)f\left(\frac{a+b}{2}\right).$$

Larger values of n yield formulas similar to Simpson's rule and the trapezoidal rule.

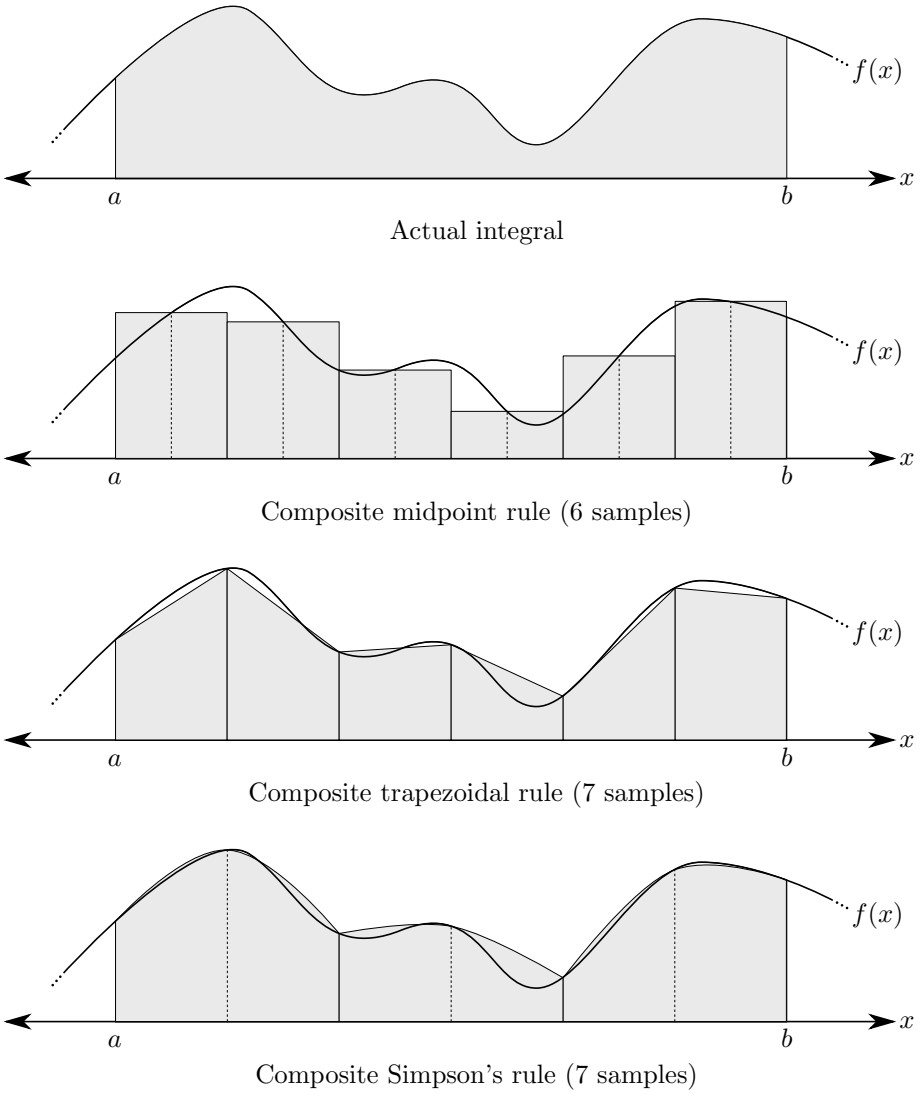


FIGURE 13.3 Composite Newton-Cotes quadrature rules; each rule is marked with the number of samples $(x_i, f(x_i))$ used to approximate the integral over six subintervals.

Composite integration. To achieve higher accuracy we likely wish to integrate $f(x)$ with more than one, two, or three sample points x_i. To do so, we can construct a composite rule out of the midpoint or trapezoidal rules above, as illustrated in Figure 13.3, by summing up smaller pieces along each interval. For example, if we subdivide $[a, b]$ into k intervals, then we can take $\Delta x \equiv \frac{b-a}{k}$ and $x_i \equiv a + (i-1)\Delta x$. Then, the composite midpoint rule is:

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{k} f\left(\frac{x_{i+1} + x_i}{2}\right)\Delta x$$

Similarly, the composite trapezoidal rule is:

$$\begin{aligned}\int_a^b f(x)\,dx &\approx \sum_{i=1}^{k}\left(\frac{f(x_i) + f(x_{i+1})}{2}\right)\Delta x \\ &= \Delta x\left(\frac{1}{2}f(a) + f(x_2) + f(x_3) + \cdots + f(x_k) + \frac{1}{2}f(b)\right) \\ &\qquad \text{after reorganizing the sum.}\end{aligned}$$

An alternative derivation of the composite midpoint rule is to apply the interpolatory quadrature formula from §13.2.1 to piecewise constant interpolation; similarly, the composite version of the trapezoidal rule comes from piecewise linear interpolation.

The composite version of Simpson's rule, also illustrated in Figure 13.3, chains together three points at a time to make parabolic approximations. Adjacent parabolas meet at every other x_i and may not share tangents. After combining terms, this quadrature rule becomes:

$$\int_a^b f(x)\,dx \approx \frac{\Delta x}{6}\left[f(a) + 4f(x_2) + 2f(x_3) + 4f(x_4) + 2f(x_5) + \cdots + 4f(x_k) + f(b)\right]$$

Accuracy. So far, we have developed a number of quadrature rules that combine the same set of $f(x_i)$'s with different weights to obtain potentially unequal approximations of the integral of f. Each approximation is based on a different interpolatory construction, so it is unclear that any of these rules is better than any other. Thus, we need to develop error estimates characterizing their respective behavior. We will use our Newton-Cotes integrators above to show how such comparisons might be carried out.

First, consider the midpoint quadrature rule on a single interval $[a, b]$. Define $c \equiv \frac{1}{2}(a+b)$. The Taylor series of f about c is:

$$f(x) = f(c) + f'(c)(x-c) + \frac{1}{2}f''(c)(x-c)^2 + \frac{1}{6}f'''(c)(x-c)^3 + \frac{1}{24}f''''(c)(x-c)^4 + \cdots$$

After integration, by symmetry about c, the odd-numbered derivatives drop out:

$$\int_a^b f(x)\,dx = (b-a)f(c) + \frac{1}{24}f''(c)(b-a)^3 + \frac{1}{1920}f''''(c)(b-a)^5 + \cdots$$

The first term of this sum is exactly the estimate of $\int_a^b f(x)\,dx$ provided by the midpoint rule, so based on this formula we can conclude that this rule is accurate up to $O(\Delta x^3)$.

Continuing, plugging a and b into our Taylor series for $f(x)$ about c shows:

$$\begin{aligned}f(a) &= f(c) + f'(c)(a-c) + \frac{1}{2}f''(c)(a-c)^2 + \frac{1}{6}f'''(c)(a-c)^3 + \cdots \\ f(b) &= f(c) + f'(c)(b-c) + \frac{1}{2}f''(c)(b-c)^2 + \frac{1}{6}f'''(c)(b-c)^3 + \cdots\end{aligned}$$

Adding these together and multiplying both sides by $\frac{1}{2}(b-a)$ shows:

$$\begin{aligned}(b-a)\frac{f(a)+f(b)}{2} &= f(c)(b-a) + \frac{1}{4}f''(c)(b-a)((a-c)^2+(b-c)^2) + \cdots \\ &= f(c)(b-a) + \frac{1}{8}f''(c)(b-a)^3 + \cdots \text{ by definition of } c\end{aligned}$$

The $f'(c)$ term vanishes for the first line by substituting $c = \frac{1}{2}(a+b)$. Now, the left hand side is the trapezoidal rule integral estimate, and the right hand side agrees with the Taylor series for $\int_a^b f(x)\,dx$ up to the cubic term. Hence, the trapezoidal rule is also $O(\Delta x^3)$ accurate in a single interval. A similar argument applies to finding an error estimate for Simpson's rule; after somewhat more involved algebra, one can show Simpson's rule has error scaling like $O(\Delta x^5)$.

We pause here to highlight a surprising result: The trapezoidal and midpoint rules have the same order of accuracy! Examining the third-order term shows that the midpoint rule is approximately two times more accurate than the trapezoidal rule, making it marginally preferable for many calculations. This observation seems counterintuitive, since the trapezoidal rule uses a linear approximation while the midpoint rule is constant. As you will see in problem 13.1, however, the midpoint rule actually recovers the integral of linear functions, explaining its extra degree of accuracy.

A notable caveat applies to this sort of analysis. Taylor's theorem only applies when Δx is sufficiently *small*; otherwise, the analysis above is meaningless. When a and b are far apart, to return to the case of small Δx, we can divide $[a, b]$ into many intervals of width Δx and apply the composite quadrature rules. The total number of intervals is $b-a/\Delta x$, so we must multiply our error estimates by $1/\Delta x$ in this case. In particular, the following orders of accuracy hold:

- Composite midpoint: $O(\Delta x^2)$
- Composite trapezoid: $O(\Delta x^2)$
- Composite Simpson: $O(\Delta x^4)$

13.2.4 Gaussian Quadrature

In some applications, we can choose the locations x_i at which f is sampled. In this case, we can optimize not only the weights for the quadrature rule but also the locations x_i to get the highest quality. This observation leads to challenging but theoretically-appealing quadrature rules, such as the Gaussian quadrature technique explored below.

The details of this technique are outside the scope of our discussion, but we provide one path to its derivation. Generalizing Example 13.7, suppose that we wish to optimize $x_1, \ldots, x_n$ *and* $w_1, \ldots, w_n$ simultaneously to increase the order of an integration scheme. Now we have $2n$ instead of n unknowns, so we can enforce equality for $2n$ examples:

$$\begin{aligned}\int_a^b f_1(x)\,dx &= w_1 f_1(x_1) + w_2 f_1(x_2) + \cdots + w_n f_1(x_n) \\ \int_a^b f_2(x)\,dx &= w_1 f_2(x_1) + w_2 f_2(x_2) + \cdots + w_n f_2(x_n) \\ &\vdots \qquad\qquad \vdots\end{aligned}$$

$$\int_a^b f_{2n}(x)\,dx = w_1 f_n(x_1) + w_2 f_n(x_2) + \cdots + w_n f_n(x_n)$$

Since both the x_i's and the w_i's are unknown, this system of equations is not linear and must be solved using more involved methods.

Example 13.8 (Gaussian quadrature)**.** If we wish to optimize weights and sample locations for polynomials on the interval $[-1, 1]$, we would have to solve the following system of polynomials [54]:

$$\begin{aligned} w_1 + w_2 &= \int_{-1}^{1} 1\,dx = 2 \\ w_1 x_1 + w_2 x_2 &= \int_{-1}^{1} x\,dx = 0 \\ w_1 x_1^2 + w_2 x_2^2 &= \int_{-1}^{1} x^2\,dx = \frac{2}{3} \\ w_1 x_1^3 + w_2 x_2^3 &= \int_{-1}^{1} x^3\,dx = 0 \end{aligned}$$

Systems like this can have multiple roots and other degeneracies that depend not only on the f_i's (typically polynomials) but also the interval over which the integral is approximated. These rules are not *progressive*, in that the x_i's chosen to integrate using n data points have little in common with those used to integrate using k data points when $k \neq n$. So, it is difficult to reuse data to achieve a better estimate with this quadrature rule. On the other hand, when they are applicable, Gaussian quadrature has the highest possible degree of accuracy for fixed n. *Kronrod* quadrature rules adapt Gaussian points to the progressive case but no longer have the highest possible order of accuracy.

13.2.5 Adaptive Quadrature

As we already have shown, there are certain functions f whose integrals are better approximated with a given quadrature rule than others; for example, the midpoint and trapezoidal rules integrate linear functions with full accuracy while sampling issues and other problems can occur if f oscillates rapidly.

Our discussion of Gaussian quadrature suggests that the placement of the x_i's can have an effect on the quality of a quadrature scheme. There still is one piece of information we have not used, however: the function values $f(x_i)$. That is, different classes or shapes of functions may require different integration methods, but so far our discussion has not taken this structure into account in any serious way.

With this situation in mind, *adaptive* quadrature strategies examine the current estimate of an integral and generate new x_i's where the integrand is undersampled. Strategies for adaptive integration often compare the output of multiple quadrature techniques, e.g. trapezoid and midpoint, with the assumption that they agree where sampling of f is sufficient, as illustrated in Figure 13.4. If they do not agree with some tolerance on a given interval, an additional sample point is generated and the integral estimates are updated.

Figure 13.5 outlines one common technique for adaptive quadrature via bisection. The idea here is to recursively subdivide intervals in which the integral estimate appears to be inaccurate. In practice, such a method must be accompanied with special consideration

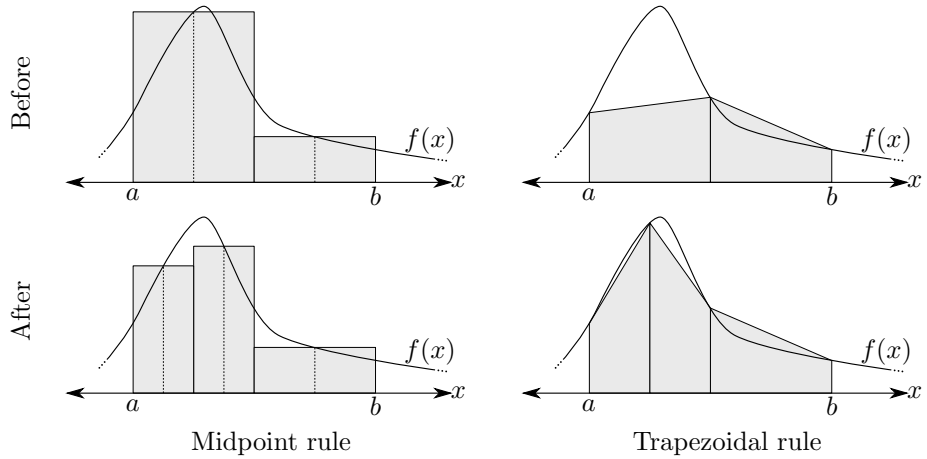


FIGURE 13.4 The trapezoidal and midpoint rules disagree considerably on the left subinterval (top), so adaptive quadrature methods subdivide in that region to get better accuracy (bottom).

```
function RECURSIVE-QUADRATURE(f(x), a, b, ε0)
  I ← QUADRATURE-RULE(f(x), a, b)
  E ← ERROR-ESTIMATE(f(x), I, a, b)
  if E < ε0 then
    return I
  else
    c ← ½(a + b)
    I1 ← RECURSIVE-QUADRATURE(f(x), a, c, ε0)
    I2 ← RECURSIVE-QUADRATURE(f(x), c, b, ε0)
    return I1 + I2
```

FIGURE 13.5 An outline for recursive quadrature via bisection. This method can use any of the quadrature rules discussed in this chapter; error estimates can be constructed e.g. by evaluating the difference between using different quadrature rules for the same interval. The parameter ε_0 is a tolerance for the quality of the quadrature rule.

when the level of recursion is too deep, helping account for the case of a function $f(x)$ that is noisy even at tiny scale.

13.2.6 Multiple Variables

Many times we wish to integrate functions $f(\vec{x})$ where $\vec{x} \in \mathbb{R}^n$. For example, when $n = 2$ we might integrate over a rectangle by computing

$$\int_a^b \int_c^d f(x, y)\, dx\, dy.$$

More generally, we might wish to find an integral $\int_\Omega f(\vec{x})\, d\vec{x}$, where Ω is some subset of $\mathbb{R}^n$.

A "curse of dimensionality" makes integration exponentially more difficult as the dimension increases. In particular, the number of sample locations $\vec{x}_i$ of $f(\vec{x})$ needed to achieve comparable quadrature accuracy for an integral in $\mathbb{R}^n$ increases exponentially in n. This observation may be disheartening but is somewhat reasonable: The more input dimensions for f, the more samples are needed to understand its behavior in all dimensions.

One way to extend single-variable integration to $\mathbb{R}^k$ is via the *iterated integral*. For example, if $f(x, y)$ is a function of two variables, suppose we wish to find $\int_a^b \int_c^d f(x, y)\, dx\, dy$. For fixed y, we can approximate the inner integral over x using a one-dimensional quadrature rule; then, we integrate these values over y using another quadrature rule. Both integration schemes induce some error, so we may need to sample $\vec{x}_i$'s more densely than in one dimension to achieve desired output quality.

Alternatively, just as we subdivided $[a, b]$ into intervals, we can subdivide Ω into triangles and rectangles in 2D, polyhedra or boxes in 3D, and so on and use interpolatory quadrature rules in each piece. For instance, one popular option is to integrate barycentric interpolants (§12.2.2), since this integral is known in closed form.

When n is high, however, it is not practical to divide the domain as suggested. In this case, we can use the randomized *Monte Carlo method*. In the most basic version of this method, we generate k random points $\vec{x}_i \in \Omega$ with uniform probability. Averaging the values $f(\vec{x}_i)$ and scaling the result by the volume $|\Omega|$ of Ω yields an approximation of $\int_\Omega f(\vec{x})\, d\vec{x}$:

$$\int_\Omega f(\vec{x})\, d\vec{x} \approx \frac{|\Omega|}{k} \sum_{i=1}^{k} f(\vec{x}_i).$$

This approximation converges like $1/\sqrt{k}$ as more sample points are added—independent of the dimension of Ω! So, in large dimensions the Monte Carlo estimate is preferable to the deterministic quadrature methods above. A proof of convergence requires some notions from probability theory, so we refer the reader to [98] or a similar reference for discussion.

One advantage of Monte Carlo techniques is that they are easily implemented and extended. Figure 13.6 provides a pseudocode implementation of Monte Carlo integration over a region $\Omega \subseteq [a, b]^n$. Even if we do not have a method for producing uniform samples in Ω directly, the more general integral can be carried out by sampling in the box $[a, b]^n$ and rejecting those samples outside Ω. This sampling is inappropriate when Ω is small relative to the bounding box $[a, b]^n$, since the odds of randomly drawing a point in Ω decrease in this case. To improve conditioning of this case, more advanced techniques *bias* their samples of $[a, b]^n$ based on evidence of where Ω takes the most space and where $f(\vec{x})$ is nontrivial.

```
function MONTE-CARLO-INTEGRAL(f(x⃗), Ω ⊆ [a,b]^n, p)
  c ← 0                                   ▷ Number of points inside Ω
  d ← 0                                   ▷ Average value
  for k ← 1, 2, ..., p                    ▷ Sample p points
    x⃗ ← UNIFORM-RANDOM([a,b]^n)
    if INSIDE(x⃗, Ω) then                  ▷ Otherwise reject
      c ← c + 1
      d ← d + f(x⃗)
  v ← (c/p)(b − a)^n                      ▷ Estimate of |Ω|
  y ← d/c                                 ▷ Average observed f(x⃗)
return vy
```

FIGURE 13.6 Pseudocode for Monte Carlo integration of a function $f(\vec{x}) : \Omega \to \mathbb{R}$.

13.2.7 Conditioning

So far we have evaluated the quality of a quadrature method by bounding its accuracy like $O(\Delta x^k)$ for small Δx; by this metric a set of quadrature weights with large k is preferable. Another measure discussed in [54] and elsewhere, however, balances out the accuracy measurements obtained using Taylor arguments by considering the *stability* of a quadrature method under perturbations of the function being integrated.

Consider the quadrature rule $Q[f] \equiv \sum_i w_i f(x_i)$. Suppose we perturb f to some other $\hat{f}$. Define $\|f - \hat{f}\|_\infty \equiv \max_{x \in [a,b]} |f(x) - \hat{f}(x)|$. Then,

$$\begin{aligned}
\frac{|Q[f] - Q[\hat{f}]|}{\|f - \hat{f}\|_\infty} &= \frac{|\sum_i w_i (f(x_i) - \hat{f}(x_i))|}{\|f - \hat{f}\|_\infty} \\
&\leq \frac{\sum_i |w_i||f(x_i) - \hat{f}(x_i)|}{\|f - \hat{f}\|_\infty} \text{ by the triangle inequality} \\
&\leq \|\vec{w}\|_\infty \text{ since } |f(x_i) - \hat{f}(x_i)| \leq \|f - \hat{f}\|_\infty \text{ by definition.}
\end{aligned}$$

So, according to this bound, the most *stable* quadrature rules are those with relatively small weights $\vec{w}$.

If we increase the order of quadrature accuracy by increasing the degree of the polynomial used in Newton-Cotes quadrature, the conditioning bound $\|\vec{w}\|$ generally becomes less favorable. In particularly degenerate circumstances, the w_i's even can take very *negative* values, echoing the degeneracies of high-order polynomial interpolation. Thus, in practice we usually prefer composite quadrature rules summing simple estimates from many small subintervals to quadrature from higher-order interpolants, which can be unstable under numerical perturbation.

13.3 DIFFERENTIATION

Numerical integration is a relatively stable problem, in that the influence of any single value $f(x)$ on $\int_a^b f(x)\,dx$ shrinks to zero as a and b become far apart. Approximating the derivative of a function $f'(x)$, on the other hand, has no such property. From the Fourier analysis perspective, one can show that the integral $\int f(x)\,dx$ generally has lower frequencies than f, while differentiating to produce f' amplifies the frequency content of f, making sampling constraints, conditioning, and stability particularly challenging for approximating f'.

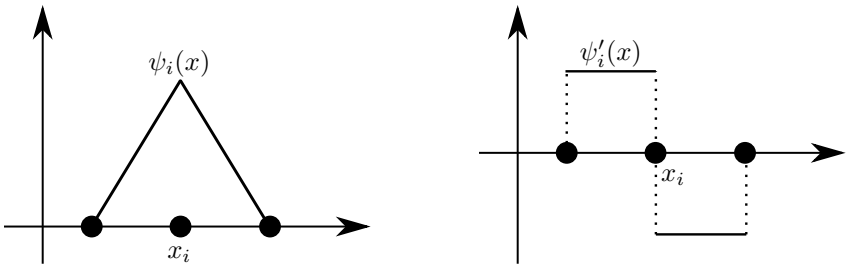


FIGURE 13.7 If a function is written in the basis of piecewise-linear "hat" functions $\psi_i(x)$, then its derivative can be written in the basis of piecewise constant functions $\psi_i'(x)$.

Despite the challenging circumstances, approximations of derivatives usually are relatively easy to implement and can be stable under sufficient smoothness assumptions. For example, while developing the secant rule, Broyden's method, and so on we used approximations of derivatives and gradients to help guide optimization routines with success on a variety of objectives.

Here we will focus on approximating f' for $f : \mathbb{R} \to \mathbb{R}$. Finding gradients and Jacobians often is accomplished by differentiating in one dimension at a time, effectively reducing to the one-dimensional problem we consider here.

13.3.1 Differentiating Basis Functions

From a mathematical perspective, perhaps the simplest use case for numerical differentiation involves functions that are constructed using interpolation routines. Just as in §13.2.1, if we can write $f(x) = \sum_i a_i \phi_i(x)$ then by linearity we know

$$f'(x) = \sum_i a_i \phi_i'(x).$$

In other words, we can think of the functions ϕ_i' as a *basis* for derivatives of functions written in the ϕ_i basis!

This phenomenon often connects different interpolatory schemes, as in Figure 13.7. For example, piecewise linear functions have piecewise constant derivatives, polynomial functions have polynomial derivatives of lower degree, and so on; in future chapters, we will see that this structure strongly influences discretizations of differential equations.

13.3.2 Finite Differences

A more common situation is that we have a function $f(x)$ that we can query but whose derivatives are unknown. This often happens when f takes on a complex form or when a user provides $f(x)$ as a subroutine without analytical information about its structure.

The definition of the derivative suggests a reasonable approximation:

$$f'(x) \equiv \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

As we might expect, for a finite $h > 0$ with small $|h|$ the expression in the limit provides an approximation of $f'(x)$.

To substantiate this intuition, we can use Taylor series to write:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \cdots$$

Rearranging this expression shows:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h)$$

Thus, the following *forward difference approximation* of f' has linear convergence:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

Similarly, flipping the sign of h shows that *backward differences* also have linear convergence:

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}$$

We actually can improve the convergence of our approximation by combining the forward and backward estimates. By Taylor's theorem we can write:

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{6}f'''(x)h^3 + \cdots \\ f(x-h) &= f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 - \frac{1}{6}f'''(x)h^3 + \cdots \\ \implies f(x+h) - f(x-h) &= 2f'(x)h + \frac{1}{3}f'''(x)h^3 + \cdots \\ \implies \frac{f(x+h) - f(x-h)}{2h} &= f'(x) + O(h^2) \end{aligned}$$

Hence, this *centered difference* gives an approximation of $f'(x)$ with quadratic convergence; this is the highest order of convergence we can expect to achieve with a divided difference. We can, however, achieve more accuracy by evaluating f at other points, e.g. $x+2h$, at the cost of additional computation time, as explored in §13.3.3.

Estimations of higher-order derivatives can be derived via similar constructions. For example, if we add together the Taylor expansions of $f(x+h)$ and $f(x-h)$ we have

$$\begin{aligned} f(x+h) + f(x-h) &= 2f(x) + f''(x)h^2 + O(h^3) \\ \implies \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} &= f''(x) + O(h^2) \end{aligned}$$

To predict similar combinations for higher derivatives, one trick is to notice that our second derivative formula can be factored differently:

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = \frac{\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}}{h}$$

That is, our approximation of the second derivative is a "finite difference of finite differences." One way to interpret this formula is shown in Figure 13.8. When we compute the forward difference approximation of f' between x and $x+h$, we can think of this slope as living at $x + {}^{h}/_{2}$; we similarly can use backward differences to place a slope at $x - {}^{h}/_{2}$. Finding the slope between these values puts the approximation back on x.

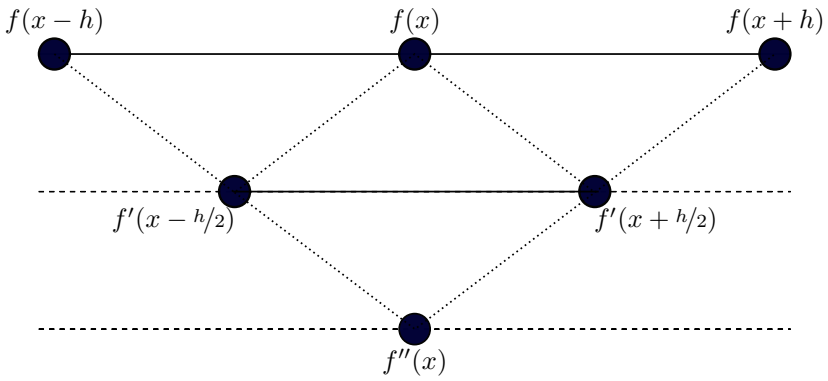


FIGURE 13.8 Computing the second derivative $f''(x)$ by divided differences can be thought of as applying the same divided difference rule once to approximate f' and a second time to approximate f''.

13.3.3 Richardson Extrapolation

One way to improve convergence of the approximations above is *Richardson extrapolation*. As an example of a more general pattern, suppose we wish to use forward differences to approximate f'. Define

$$D(h) \equiv \frac{f(x+h) - f(x)}{h}.$$

We have argued that $D(h)$ approaches $f'(x)$ as $h \to 0$. More specifically, however, from our discussion in §13.3.2 we know that $D(h)$ takes the form:

$$D(h) = f'(x) + \frac{1}{2}f''(x)h + O(h^2)$$

Suppose we know $D(h)$ and $D(\alpha h)$ for some $0 < \alpha < 1$. We know:

$$D(\alpha h) = f'(x) + \frac{1}{2}f''(x)\alpha h + O(h^2)$$

We can write these two relationships in a matrix:

$$\begin{pmatrix} 1 & \frac{1}{2}h \\ 1 & \frac{1}{2}\alpha h \end{pmatrix} \begin{pmatrix} f'(x) \\ f''(x) \end{pmatrix} = \begin{pmatrix} D(h) \\ D(\alpha h) \end{pmatrix} + O(h^2)$$

Or equivalently,

$$\begin{pmatrix} f'(x) \\ f''(x) \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2}h \\ 1 & \frac{1}{2}\alpha h \end{pmatrix}^{-1} \begin{pmatrix} D(h) \\ D(\alpha h) \end{pmatrix} + O(h^2)$$

That is, we took an $O(h)$ approximation of $f'(x)$ using $D(h)$ and made it into an $O(h^2)$ approximation! This clever technique is a method for *sequence acceleration*, since it improves the order of convergence of the approximation $D(h)$.

The same method is applicable more generally to many other problems including numerical integration. Richardson extrapolation even can be applied recursively to make higher and higher order approximations of the same quantity.

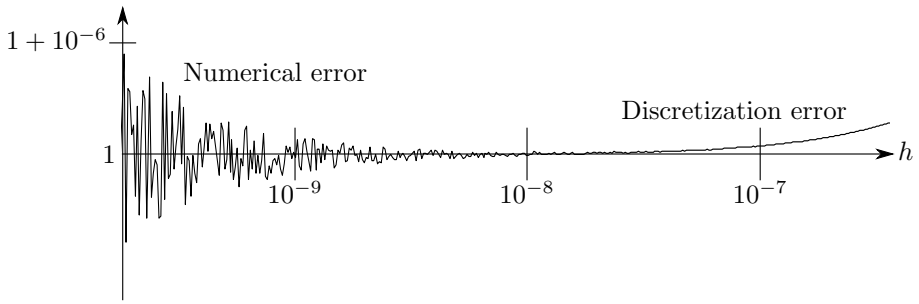


FIGURE 13.9 The finite difference derivative $1/h(f(x+h)-f(x))$ as a function of h for $f(x) = x^2/2$, computed using IEEE floating-point arithmetic; when h is too small, the approximation suffers from numerical issues, while large h yields discretization error. The horizontal axis is on a logarithmic scale, and the vertical axis scales linearly.

13.3.4 Choosing the Step Size

Unlike quadrature, numerical differentiation has a curious property. It appears that any formula above can be arbitrarily accurate without extra computational cost by choosing a sufficiently small h. This observation is appealing from the perspective that we can achieve higher-quality approximations without additional computation time.

The catch, however, is that we must divide by h; the smaller the value of h, the more similar the values $f(x)$ and $f(x + h)$, and in finite-precision arithmetic, adding and/or dividing by near-zero values induces numerical issues and instabilities. Thus, there is a range of h values that are not large enough to induce significant discretization error and not small enough to make for numerical problems; Figure 13.9 shows an example for differentiating a simple function in IEEE floating point arithmetic.

Similarly, suppose as in §13.2.7 that due to noise rather than evaluating $f(x)$ we actually receive perturbed values from a function $\hat{f}(x)$ satisfying $\|f - \hat{f}\|_\infty \leq \varepsilon$. Then, we might bound the error of computing this difference quotient:

$$\begin{aligned}
\left|\frac{\hat{f}(x+h)-\hat{f}(x)}{h} - f'(x)\right| &\leq \left|\frac{\hat{f}(x+h)-\hat{f}(x)}{h} - \frac{f(x+h)-f(x)}{h}\right| + O(h) \\
&\qquad \text{by our previous bound} \\
&\leq \left|\frac{(\hat{f}(x+h)-f(x+h)) - (\hat{f}(x)-f(x))}{h}\right| + O(h) \\
&\leq \frac{2\varepsilon}{h} + O(h) \text{ since } \|f-\hat{f}\|_\infty \leq \varepsilon
\end{aligned}$$

For fixed $\varepsilon > 0$, this bound actually can get worse if we take $h \to 0$; instead, we should choose h to balance the $2\varepsilon/h$ and $O(h)$ terms to get minimal error. That is, if we cannot compute values of $f(x)$ exactly, taking larger $h > 0$ can actually *improve* the quality of the estimate of $f'(x)$. Problem 13.6f has a similar conclusion about a method for numerical integration.

13.3.5 Automatic Differentiation

As we have seen, typical algorithms for numerical differentiation are relatively fast since they involve little more than computing a difference quotient. Their main drawback is numerical, in that finite-precision arithmetic and/or inexact evaluation of functions fundamentally limit the quality of the output. Noisy or rapidly-varying functions are thus difficult to differentiate numerically with any confidence.

For some applications, on the other end of the spectrum between computational efficiency and numerical quality lies the technique of *automatic differentiation* ("autodiff"), which is not subject to any discretization error [8]. These techniques take advantage of the chain rule and other properties derivatives to compute derivatives exactly.

"Forward" automatic differentiation is particularly straightforward to implement. Suppose we have two variables u and v, stored using floating point values. We can store alongside these variables additional values $u' \equiv {}^{du}/_{dt}$ and $v' \equiv {}^{dv}/_{dt}$ for some independent variable t; we can define a new data type holding two values $[u, u']$ and $[v, v']$. We can define an algebra on these pairs that encodes typical operations:

$$\begin{aligned}
[u, u'] + [v, v'] &\equiv [u + v, u' + v'] \\
c[u, u'] &\equiv [cu, cu'] \\
[u, u'] \cdot [v, v'] &\equiv [uv, uv' + u'v] \\
[u, u'] \div [v, v'] &\equiv \left[\frac{u}{v}, \frac{vu' - uv'}{v^2}\right] \\
\exp([u, u']) &\equiv [e^u, u'e^u] \\
\ln([u, u']) &\equiv \left[\ln u, \frac{u'}{u}\right] \\
\cos([u, u']) &\equiv [\cos u, -u' \sin u] \\
\vdots \quad & \quad \vdots
\end{aligned}$$

Starting with the pair $t \equiv [t_0, 1]$—since ${}^{dt}/_{dt} = 1$—we can build up a function $f(t)$ and its derivative $f'(t)$ simultaneously using these rules. If they are implemented in a programming language supporting operator overloading, the additional derivative computations can be completely transparent to the implementer.

The method we just describes builds up the derivative $f'(t)$ in parallel with building $y = f(t)$. "Backward" automatic differentiation is an alternative algorithm that can require fewer function evaluations in exchange for more memory usage and more complex implementation. This technique builds up a graph representing the steps of computing $f(t)$ as a sequence of elementary operations. Then, rather than starting from the fact ${}^{dt}/_{dt} = 1$ and working forward to ${}^{dy}/_{dt}$, backward automatic differentiation starts with ${}^{dy}/_{dy} = 1$ and works backward from the same rules. Backward automatic differentiation can avoid unnecessary computations particularly when y is a function of multiple variables. For instance, suppose we can write $f(t_1, t_2) = f_1(t_1) + f_2(t_2)$; in this case, backward automatic differentiation does not need to differentiate f_1 with respect to t_2 or f_2 with respect to t_1.

Automatic differentiation is widely regarded as an under-appreciated numerical technique, in that it yields *exact* derivatives of functions with limited additional implementation effort. This property can be particularly valuable when prototyping methods making use of optimization methods requiring derivatives or Hessians, since it can be very time-consuming to work out the derivatives of potential options while engineering the right objective function for a given problem. The cost of this convenience, however, is computational efficiency,

since in effect automatic differentiation methods do not simplify expressions for derivatives but rather apply the most obvious rules.

13.3.6 Integrated Quantities and Structure Preservation

Continuing in our consideration of alternatives to numerical differentiation, we outline an approach that has gained popularity in the geometry and computer graphics communities for dealing with curvature and other differential measures of shape.

As we have seen, a typical pattern from numerical analysis is to prove that properties of approximated derivatives hold as $\Delta x \to 0$ for some measure of spacing Δx. While this type of analysis provides intuition relating discrete computations to continuous notions from calculus, it neglects a key fact: In reality, we must fix $\Delta x > 0$. Understanding what happens in the $\Delta x > 0$ regime can be equally important to the $\Delta x \to 0$ limit, especially when taking coarse approximations. For example, in computational geometry, it may be desirable to link measures like curvature of smooth shape directly to discrete values like lengths and angles that can be computed on complexes of polygons.

With this new view, some techniques involving derivatives, integrals, and other quantities are designed with *structure preservation* in mind, yielding "discrete" rather than "discretized" analogs of continuous quantities [49]. That is, rather than asking that structure from continuous calculus emerges as $\Delta x \to 0$, we design differentiators and integrators for which certain theorems from continuous mathematics hold exactly.

One central technique in this domain is the use of *integrated quantities* to encode derivatives. As a basic example, suppose we are sampling $f(t)$ and have computed $f(t_1), f(t_2), \ldots, f(t_k)$ for some discrete set of times $t_1 < t_2 < \cdots < t_k$. Rather than using divided differences to approximate the derivative f', we can use the Fundamental Theorem of Calculus to show:

$$\int_{t_i}^{t_{i+1}} f'(t)\,dt = f(t_{i+1}) - f(t_i)$$

This formula may not appear remarkable beyond first-year calculus, but it encodes a deep idea. The difference $f(t_{i+1}) - f(t_i)$ on the right side is computable *exactly* from the samples $f(t_1), f(t_2), \ldots, f(t_k)$, while the quantity on the left is an averaged version of the derivative f'. By substituting integrated versions of f' into computations whenever possible, we can carry out discrete analogs of continuous calculus for which certain theorems and properties hold exactly rather than in the limit.

Example 13.9 (Curvature of a 2D curve, [49])**.** In the continuous theory of differential geometry, a curve Γ on the two-dimensional plane can be parameterized as a function $\gamma(s) : \mathbb{R} \to \mathbb{R}^2$ satisfying $\gamma'(s) \neq \vec{0}$ for all s. Assume that $\|\gamma'(s)\|_2 = 1$ for all s; such an arc length parameterization is always possible by moving along the curve with constant speed. Then, Γ has unit tangent vector $\vec{T}(s) \equiv \gamma'(s)$. If we write $\vec{T}(s) \equiv (\cos\theta(s), \sin\theta(s))$ for angle $\theta(s)$, then the *curvature* of $\gamma(s)$ is given by the derivative $\kappa(s) \equiv \theta'(s)$. This notation is illustrated in Figure 13.10 alongside notation for the discretization below.

Suppose Γ is closed, that is, $\gamma(s_0) = \gamma(s_1)$ for some $s_0, s_1 \in \mathbb{R}$. Then, the *turning number theorem* from topology states:

$$\int_{s_0}^{s_1} \kappa(s)\,ds = 2\pi k$$

for some integer k. Intuitively, this theorem represents the fact that $\vec{T}(s_0) = \vec{T}(s_1)$, and hence θ took some number of loops around the full circle.

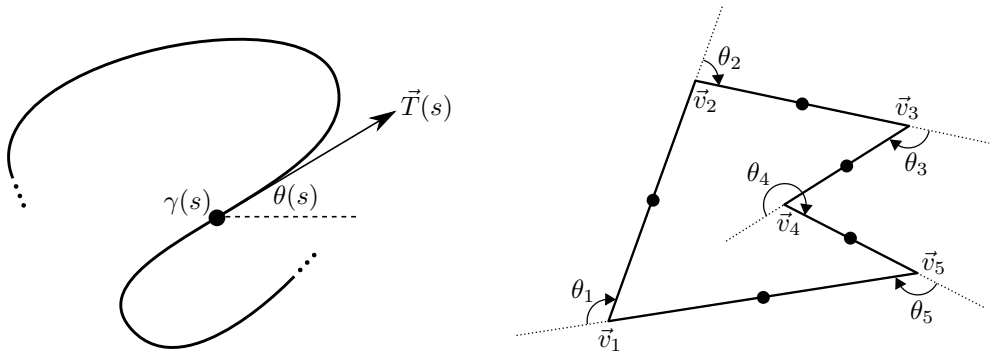


FIGURE 13.10 Notation for Example 13.9; each curve segment Γ_i is the union of the two half-segments adjacent to $\vec{v}_i$, bounded by the marked midpoints.

A typical discretization of a two-dimensional curve is as a sequence of line segments $\vec{v}_i \leftrightarrow \vec{v}_{i+1}$. Approximating $\kappa(s)$ on such a curve can be a challenging problem, since κ is related to the second derivative γ''. Instead, suppose at each joint $\vec{v}_i$ we define the *integrated curvature* over the two half-segments around $\vec{v}_i$ to be the turning angle θ_i given by the π minus the angle between the two segments adjacent to $\vec{v}_i$.

Partition the discretization of Γ into pairs of half-segments Γ_i. Then, if Γ is closed,

$$\begin{aligned}\int_\Gamma \kappa\, ds &= \sum_i \int_{\Gamma_i} \kappa\, ds \text{ by breaking into individual terms} \\ &= \sum_i \theta_i \text{ by definition of integrated curvature} \\ &= 2\pi k,\end{aligned}$$

where the final equality comes from the fact that the discrete Γ is a polygon, and we are summing its exterior angles. That is, for our choice of *discrete curvature*, the turning number theorem holds *exactly* even for coarse approximations of Γ rather than becoming closer and closer to true as the lengths $|\Gamma_i| \to 0$. In this sense, the integrated turning-angle curvature has more properties in common with the continuous curvature of a curve $\gamma(s)$ than an inexact but convergent discretization coming from divided differences.

Our example above shows a typical structure-preserving treatment of a derivative quantity, in this case the curvature of a two-dimensional curve, accompanied by a discrete structure—the turning number theorem—holding without taking any limit as $\Delta x \to 0$. We have not shown, however, that the value θ_i—or more precisely some non-integrated pointwise approximation like $\theta_i/|\Gamma_i|$—actually converges to the curvature of Γ. This type of convergence does *not* always hold, and in some cases it is impossible to preserve structure exactly and converge as $\Delta x \to 0$ simultaneously [122]; such convergence issues are the topic of active research at the intersection of numerical methods and geometry processing.

13.4 EXERCISES

13.1 Show that the midpoint rule is exact for the function $f(x) = mx + c$ along any interval $x \in [a, b]$.

13.2 Derive α, β, and x_1 such that the following quadrature rule holds exactly for polynomials of degree ≤ 2 :

$$\int_0^2 f(x)\,dx \approx \alpha f(0) + \beta f(x_1)$$

13.3 Suppose we are given a quadrature rule of the form $\int_0^1 f(x)\,dx \approx af(0) + bf(1)$ for some $a, b \in \mathbb{R}$. Propose a composite rule for approximating $\int_0^1 f(x)\,dx$ given $n+1$ closed sample points $y_0 \equiv f(0), y_1 \equiv f(1/n), y_2 \equiv f(2/n), \ldots, y_n \equiv f(1)$.

13.4 Some quadrature problems can be solved by applying a suitable change of variables:

(a) Our strategies for quadrature break down when the interval of integration is not of finite length. Derive the following relationships for $f : \mathbb{R} \to \mathbb{R}$:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-1}^{1} f\left(\frac{t}{1-t^2}\right)\frac{1+t^2}{(1-t^2)^2}\,dt$$
$$\int_0^{\infty} f(x)\,dx = \int_0^1 \frac{f(-\ln t)}{t}\,dt$$
$$\int_c^{\infty} f(x)\,dx = \int_0^1 f\left(c + \frac{t}{1-t}\right)\cdot\frac{1}{(1-t)^2}\,dt$$

How can these formulas be used to integrate over intervals of infinite length? What might be a drawback of evenly spacing t samples?

(b) Suppose $f : [-1, 1] \to \mathbb{R}$ can be written:

$$f(\cos\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\theta)$$

Then, show:

$$\int_{-1}^{1} f(x)\,dx = a_0 + \sum_{k=1}^{\infty} \frac{2a_{2k}}{1-(2k)^2}.$$

This formula provides a way to integrate a function given its Fourier series [23].

13.5 The methods in this chapter for differentiation were limited to single-valued functions $f : \mathbb{R} \to \mathbb{R}$. Suppose $g : \mathbb{R}^n \to \mathbb{R}^m$. How would you use these techniques to approximate the Jacobian Dg? How does the timing of your approach scale with m and n?

13.6 ("Lanczos differentiator," [72]) Suppose $f(t)$ is a smooth function.

(a) Suppose we sample $f(t)$ at $t = kh$ for $k \in \{-n, -n+1, \ldots, 0, \ldots, n\}$, yielding samples $y_{-n} = f(-nh), y_{-n+1} = f((-n+1)h), \ldots, y_n = f(nh)$. Show that the parabola $p(t) = at^2 + bt + c$ optimally fitting these data points via least-squares satisfies

$$p'(0) = \frac{\sum_k k y_k}{h\sum_k k^2}.$$

(b) Use this formula to propose approximations of $f'(0)$ when $n = 1, 2, 3$.

(c) By taking a limit as $h \to 0$, motivate the following formula for "differentiation by integration:"

$$f'(0) = \lim_{h\to 0} \frac{3}{2h^3} \int_{-h}^{h} tf(t)\,dt.$$

This formula provides one connection between numerical methods for integration and differentiation.

(d) Show that when $h > 0$,

$$\frac{3}{2h^3} \int_{-h}^{h} tf(t)\,dt = f'(0) + O(h^2)$$

(e) Denote $D_h f \equiv \frac{3}{2h^3} \int_{-h}^{h} tf(t)\,dt$. Suppose thanks to noise we actually observe $f^\varepsilon(t)$ satisfying $|f(t) - f^\varepsilon(t)| \leq \varepsilon$ for all t. Show the following relationship:

$$|D_h f^\varepsilon(0) - f'(0)| \leq \frac{3\varepsilon}{2h} + O(h^2)$$

(f) Suppose the second term in part 13.6e is bounded above by $\frac{M}{10}h^2$; this is the case when $|f'''(t)| \leq M$ everywhere [50]. Show that with the right choice of h, the integral approximation from part 13.6e is within $O(\varepsilon^{2/3})$ of $f'(0)$.
Note: Your choice of h effectively trades off between numerical approximation error from using the "differentiation by integration" formula and noise approximating f with f^ε. This property makes the Lanczos approximation effective for certain noisy functions.

13.7 Propose an extension of forward automatic differentiation to maintaining first and second derivatives in triplets $[u, u', u'']$. Provide analogous formulas for the operations listed in §13.3.5 given $[u, u', u'']$ and $[v, v', v'']$.

13.8 The problem of numerical differentiation is challenging for functions that are noisy or contain high frequencies. One way to stabilize such a calculation is to consider multiple samples simultaneously [1]. For this problem, assume $f : [0, 1] \to \mathbb{R}$ is differentiable.

(a) By the Fundamental Theorem of Calculus, there exists $c \in \mathbb{R}$ such that

$$f(x) = c + \int_0^x f'(\bar{x})\,d\bar{x}.$$

Suppose we sample $f(x)$ at evenly-spaced points $x_0 = 0, x_1 = h, x_2 = 2h, \ldots, x_n = 1$ and wish to approximate the first derivative $f'(x)$ at $x_1 - h/2, x_2 - h/2, \ldots, x_n - h/2$. If we label our samples of $f'(x)$ as $a_1, \ldots, a_n$, write a least-squares problem in the a_i's and an additional unknown c approximating this integral relationship.

(b) Propose a Tikhonov regularizer for this problem.

(c) We also could have written

$$f(x) = \tilde{c} - \int_x^1 f'(\bar{x})\,d\bar{x}.$$

Does your approximation of $f'(\bar{x})$ change if you use this formula?

13.9 Richardson extrapolation (§13.3.3) can also be applied to quadrature to derive the *Romberg* quadrature rules. Here we will derive Romberg integration for $f : [a, b] \to \mathbb{R}$.

(a) Suppose we divide $[a, b]$ into 2^k subintervals for $k \geq 0$. Denote by $T_{k,0}$ the result of applying the composite trapezoidal rule to $f(x)$ to this subdivision. Show that there exists a constant C dependent on f but not k such that:

$$\int_a^b f(x)\,dx = T_{k,0} + Ch^2 + O(h^4),$$

where $h(k) = {}^{(b-a)}/_{2^k}$.

(b) Use Richardson extrapolation to derive an estimate $T_{k,1}$ of the integral that is accurate up to $O(h^4)$.
Hint: Combine the $T_{k,0}$'s.

(c) Assume that the error expansion for the trapezoidal rule continues in a similar fashion:

$$\int_a^b f(x)\,dx = T_{k,0} + C_2h^2 + C_4h^4 + C_6h^6 + \cdots .$$

By iteratively applying Richardson extrapolation, propose values $T_{k,j}$ for $j \leq k$ that can be used to achieve arbitrarily high-order estimates of the desired integral.
Hint: You should be able to define $T_{k,j}$ as a linear combination of $T_{k,j-1}$ and $T_{k-1,j-1}$.

13.10 Give examples of closed and open Newton-Cotes quadrature rules with negative coefficients for integrating $f(x)$ on $[0, 1]$. What unnatural properties can be exhibited by these approximations?

13.11 Provide a sequence of differentiable functions $f_k : [0, 1] \to \mathbb{R}$ and a function $f : [0, 1] \to \mathbb{R}$ such that $\max_{x\in[0,1]} |f_k(x)-f(x)| \to 0$ as $k \to \infty$ but $\max_{x\in[0,1]} |f'_k(x)-f'(x)| \to \infty$. What does this example imply about numerical differentiation when function values are noisy? Is a similar counterexample possible for integration when f and the f_k's are differentiable?

CHAPTER 14

Ordinary Differential Equations

CONTENTS

14.1 Motivation 294
14.2 Theory of ODEs 295
14.2.1 Basic Notions 295
14.2.2 Existence and Uniqueness 299
14.2.3 Model Equations 299
14.3 Time-Stepping Schemes 301
14.3.1 Forward Euler 301
14.3.2 Backward Euler 303
14.3.3 Trapezoidal Method 304
14.3.4 Runge-Kutta Methods 305
14.3.5 Exponential Integrators 307
14.4 Multivalue Methods 308
14.4.1 Newmark Schemes 308
14.4.2 Staggered Grid and Leapfrog 311
14.5 Comparison of Integrators 312

WE motivated the problem of interpolation in Chapter 12 by transitioning from *analyzing* functions to *finding* functions. That is, in problems like interpolation and regression, the unknown is a function f, and the job of the algorithm is to fill in $f(\vec{x})$ at positions $\vec{x}$ where it is unknown.

In this chapter and the next, our unknown will continue to be a function f, but rather than filling in missing values we would like to solve more complex design problems like the following:

- Find f approximating some other function f_0 but satisfying additional criteria (smoothness, continuity, boundedness, etc.).
- Simulate some dynamical or physical relationship as $f(t)$ where t is time.
- Find f with similar values to f_0 but certain properties in common with a different function g_0.

In each of these cases, our unknown is a function f, but our criterion for success is more involved than "matches a given set of data points."

The theories of ordinary differential equations (ODEs) and partial differential equations (PDEs) involve the case where we wish to find a function $f(\vec{x})$ based on information about

or relationships between its derivatives. We already have solved one problem in this class in our discussion of quadrature: Given $f'(t)$, quadrature can provide an approximation of $f(t)$ using integration.

In this chapter, we will consider the case of *ordinary* differential equations and in particular *initial value problems*. In these problems, the unknown is a function $f(t) : \mathbb{R} \to \mathbb{R}^n$, and we are given $f(0)$ as well as an equation satisfied by f and its derivatives. Our goal is to predict $f(t)$ for $t > 0$. We will provide examples of ODEs appearing in practice and then will proceed to describe common solution techniques.

14.1 MOTIVATION

ODEs appear in nearly every branch of science, and hence it is not difficult to identify target applications of solution techniques. We choose a few representative examples both from the computational and scientific literatures:

Example 14.1 (Newton's Second Law)**.** Continuing from §5.1.2, recall that Newton's Second Law of Motion states $\vec{F} = m\vec{a}$, that is, the total force on an object is equal to its mass times its acceleration. If we simulate n particles simultaneously as they move in three-dimensional space, we can think of combining all their positions into a single vector $\vec{x}(t) \in \mathbb{R}^{3n}$. Similarly, we can write a function $\vec{F}(t, \vec{x}, \vec{x}') \in \mathbb{R}^{3n}$ taking the current time, the positions of the particles, and their velocities and returning the total force on each particle divided by its mass. This function can take into account interrelationships between particles (e.g. gravitational forces, springs, or intermolecular bonds), external effects like wind resistance (which depends on $\vec{x}'$), external forces varying with time t, and so on. Then, to find the positions of all the particles as functions of time, we can integrate Newton's second law forward in time by solving the equation $\vec{x}'' = \vec{F}(t, \vec{x}, \vec{x}')$. We usually are given the positions and velocities of all the particles at time $t = 0$ as a starting condition.

Example 14.2 (Protein folding)**.** On a small scale, the equations governing motions of molecules stem from Newton's laws or—at even smaller scales—the Schrödinger equation of quantum mechanics. One particularly challenging case is that of *protein folding*, in which the geometric structure of a protein is predicted by simulating intermolecular forces over time. These forces take many nonlinear forms that continue to challenge researchers in computational biology due in large part to a variety of *time scales*: The same forces that cause protein folding and related phenomena also can make molecules vibrate rapidly, and the disparate time scales of these two different behaviors makes them difficult to capture simultaneously.

Example 14.3 (Gradient descent)**.** Suppose we wish to minimize an energy function $E(\vec{x})$ over all $\vec{x}$. Especially if E is a convex function, the most straightforward option for minimization from Chapter 8 is gradient descent with a constant step size or "learning rate." In particular, since $-\nabla E(\vec{x})$ points in the direction along which E decreases the most from a given $\vec{x}$, we can iterate:

$$\vec{x}_{i+i} \equiv \vec{x}_i - h\nabla E(\vec{x}),$$

for fixed $h > 0$. We can rewrite this relationship as

$$\frac{\vec{x}_{i+1} - \vec{x}_i}{h} = -\nabla E(\vec{x}).$$

In the style of §13.3, we might think of $\vec{x}_k$ as a sample of a function $\vec{x}(t)$ at $t = hk$. Heuristically, taking $h \to 0$ in our formula for gradient descent would show

$$\vec{x}'(t) = -\nabla E(\vec{x}).$$

If we take $\vec{x}(0)$ to be an initial guess of the location where $E(\vec{x})$ is minimized, then this ODE is a *continuous model of gradient descent*. It can be thought of as the equation of a path smoothly walking "downhill" along the heights provided by E.

For example, suppose we wish to solve $A\vec{x} = \vec{b}$ for symmetric positive definite A. We know from §10.1.1 that this is equivalent to minimizing $E(\vec{x}) \equiv \frac{1}{2}\vec{x}^\top A\vec{x} - \vec{b}^\top\vec{x} + c$. Thus, we could attempt to solve the ODE $\vec{x}' = -\nabla f(\vec{x}) = \vec{b} - A\vec{x}$. As $t \to \infty$, we expect $\vec{x}(t)$ to better and better satisfy the linear system.

Example 14.4 (Crowd simulation)**.** Suppose we are writing video game software requiring realistic simulation of virtual crowds of humans, animals, spaceships, and the like. One way to generate plausible motion is to use differential equations. In this technique, the velocity of a member of the crowd is determined as a function of its environment; for example, in human crowds the proximity of other humans, distance to obstacles, and so on can affect the direction a given agent is moving. These rules can be simple, but in the aggregate their interaction is complex. Stable integrators for differential equations underlie this machinery, since we do not wish to have noticeably unrealistic or unphysical behavior.

14.2 THEORY OF ODES

A full treatment of the theory of ordinary differential equations is outside the scope of our discussion, and we refer the reader to [59] or any other basic text for details from this classical theory. We highlight results here that will be relevant to development in future sections.

14.2.1 Basic Notions

The most general initial value problem takes the following form:

$$\begin{array}{rl} \text{Find} & f(t) : \mathbb{R} \to \mathbb{R}^n \\ \text{satisfying} & F[t, f(t), f'(t), f''(t), \ldots, f^{(k)}(t)] = \vec{0} \\ \text{given} & f(0), f'(0), f''(0), \ldots, f^{(k-1)}(0). \end{array}$$

Here, F is some relationship between f and all its derivatives; we use $f^{(\ell)}$ to denote the ℓ-th derivative of f. The functions f and F can be *multidimensional*, taking on values in $\mathbb{R}^n$ rather than $\mathbb{R}$, but by convention and for convenience of notation we will omit the vector sign. We also will use the notation $\vec{y} \equiv f(t)$ as an alternative to writing $f(t)$ when the t dependence is clear; in this case, derivatives will be notated $\vec{y}' \equiv f'(t)$, $\vec{y}'' \equiv f''(t)$, and so on.

Example 14.5 (Canonical ODE form)**.** Suppose we wish to solve the ODE $y'' = ty' \cos y$. In the general form above, the ODE can be written $F[t, y, y', y''] = 0$, where $F[t, a, b, c] \equiv tb \cos a - c$.

We can think of ODEs as determining *evolution* of f over time t; we know f and its derivatives at time zero and wish to predict it moving forward. Given this basic definition,

ODEs can take many forms even in a single variable. For instance, denote $y = f(t)$ for $y \in \mathbb{R}^1$. Then, examples of ODEs include the following:

Example ODE	Distinguishing properties
$y' = 1 + \cos t$	Can be solved by integrating both sides; in the numerical case can be approached using quadrature
$y' = ay$	Linear in y, no dependence on time t
$y' = ay + e^t$	Time- and value-dependent
$y'' + 3y' - y = t$	Involves multiple derivatives of y
$y'' \sin y = e^{ty'}$	Nonlinear in y and t

We will restrict most of our discussion to the case of *explicit* ODEs, in which the highest-order derivative can be isolated:

Definition 14.1 (Explicit ODE)**.** An ODE is *explicit* if can be written in the form

$$f^{(k)}(t) = F[t, f(t), f'(t), f''(t), \ldots, f^{(k-1)}(t)].$$

Certain *implicit* ODEs can be converted to explicit form by solving a root-finding problem, for example using the machinery introduced in Chapter 7, but this approach can fail in the presence of multiple roots.

Generalizing the trick in §5.1.2, any explicit ODE can be converted to a first-order equation $f'(t) = F[t, f(t)]$ by adding to the dimensionality of f. This observation implies that single-derivative methods are sufficient for solving a wide class of ODEs. To see this relationship for the second-order ODE $y'' = F[t, y, y']$, recall that

$$\frac{d^2y}{dt^2} = \frac{d}{dt}\left(\frac{dy}{dt}\right).$$

If we define an intermediate variable $z \equiv {}^{dy}/_{dt}$, then we can expand to the following first-order system:

$$\frac{d}{dt}\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} z \\ F[t, y, z] \end{pmatrix}.$$

More generally, if we wish to solve the explicit problem

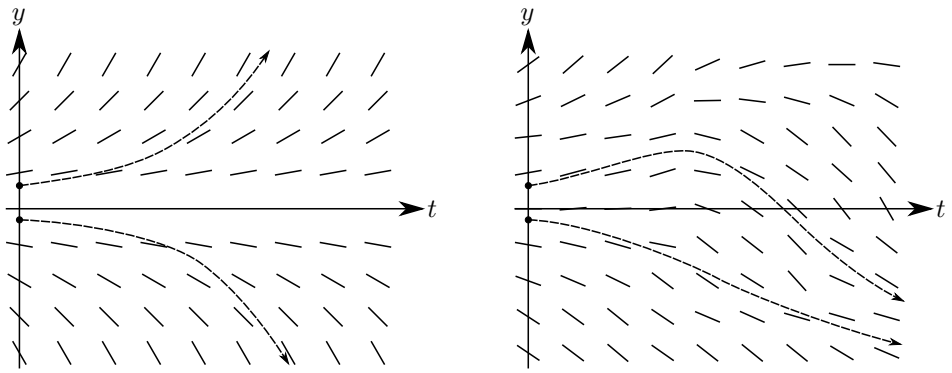
$$f^{(k)}(t) = F[t, f(t), f'(t), f''(t), \ldots, f^{(k-1)}(t)]$$

for $f : \mathbb{R} \to \mathbb{R}^n$, then instead we can solve the first-order ODE in dimension $n(k+1)$:

$$\frac{d}{dt}\begin{pmatrix} f_0(t) \\ f_1(t) \\ f_2(t) \\ \vdots \\ f_{k-1}(t) \\ f_k(t) \end{pmatrix} = \begin{pmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ \vdots \\ f_k(t) \\ F[t, f_0(t), f_1(t), \ldots, f_{k-1}(t)] \end{pmatrix}$$

Here, we denote $f_i(t) : \mathbb{R} \to \mathbb{R}^n$ as the i-th derivative of $f_0(t)$, which satisfies the original ODE. To check, our expanded system above implies $f_1(t) = f_0'(t)$, $f_2(t) = f_1'(t) = f_0''(t)$, and so on; thus, the final row encodes the original ODE.

This trick simplifies notation and allows us to emphasize first-order ODEs, but some care should be taken to understand that it does come with a cost. In particular, the expansion



Time-independent Time-dependent

FIGURE 14.1 First-order ODEs in one variable $y' = F[t, y]$ can be visualized using *slope fields* on the (t, y) plane. Here, short line segments show the slope $F[t, y]$ at each sampled point; solution curves $y(t)$ shown as dotted lines start at $(0, y(0))$ and follow the slope field as their tangents. We show an example of a time-independent ("autonomous") ODE $y' = F[y]$ and an example of a time-dependent ODE $y' = F[t, y]$.

above replaces ODEs with potentially many derivatives with ODEs containing one derivative but producing several more outputs. We will return to this trade-off between dimensionality and number of derivatives when designing methods for second-order ODEs in §14.4.2.

Example 14.6 (ODE expansion)**.** Suppose we wish to solve $y''' = 3y'' - 2y' + y$ where $y(t) : \mathbb{R} \to \mathbb{R}$. This equation is equivalent to:

$$\frac{d}{dt}\begin{pmatrix} y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -2 & 3 \end{pmatrix}\begin{pmatrix} y \\ z \\ w \end{pmatrix}$$

In the interests of making our problem as simple as possible, we can further restrict our consideration to *autonomous* ODEs. These equations are of the form $f'(t) = F[f(t)]$, that is, F has no dependence on t (or higher-order derivatives of f, as removed above). To do so, we use the fact $d/dt(t) = 1$. Hence, if we take $g(t) = t$, for the ODE $f'(t) = F[t, f(t)]$ we can write:

$$\frac{d}{dt}\begin{pmatrix} g(t) \\ f(t) \end{pmatrix} = \begin{pmatrix} 1 \\ F[g(t), f(t)] \end{pmatrix}$$

with an additional initial condition $g(0) = 0$.

It is possible to visualize the behavior and classification of low-dimensional ODEs in many ways. For instance, if the unknown $f(t)$ is a function of a single variable, then we can think of $F[f(t)]$ as providing the slope of $f(t)$, as shown in Figure 14.1. For higher-order ODEs, it can be useful to plot $f(t)$ and its derivatives, such as the pendulum example shown in Figure 14.2. In higher dimensions, it may be possible only to show typical solution paths, as in Figure 14.3.

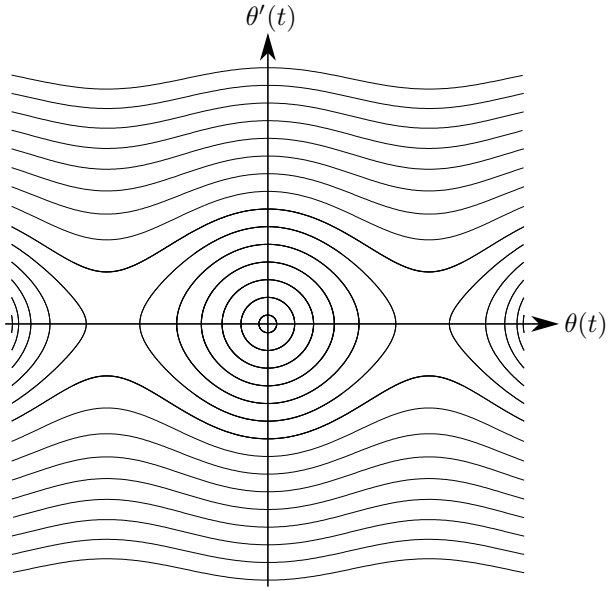


FIGURE 14.2 The *phase space diagram* of a pendulum, which satisfies the ODE $\theta'' = -\sin\theta$. Here, the horizontal axis shows position θ of the pendulum as it swings (as an angle from vertical), and the vertical axis shows the angular velocity θ'. Each path represents the motion of a pendulum with different starting conditions; the time t is not depicted. Rings indicate a swinging pendulum, while waves indicate that the pendulum is doing complete revolutions.

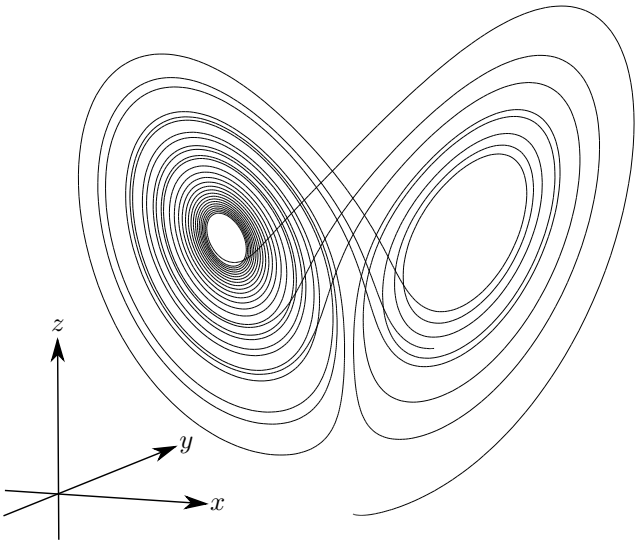


FIGURE 14.3 The *trace* of an ODE solution $(x(t), y(t), z(t))$ shows typical behavior without showing the velocity of the path or dependence on time t; here we show the a solution to the *Lorenz equations* (known as a "Lorenz attractor") $x' = \sigma(y-x), y' = x(\rho - z) - y, z' = xy - \beta z$ integrated numerically ($\rho = 28, \sigma = 10, \beta = 8/3$).

14.2.2 Existence and Uniqueness

Before we discretize the initial value ODE problem, we should acknowledge that not all differential equations are solvable. On the opposite extreme, other differential equations admit multiple solutions. Existence and uniqueness can be challenging to prove, but without these properties we cannot hold numerical methods responsible for failure to recover a reasonable solution. In some sense, numerical differential equations techniques can be thought of as filling the gap between knowing a solution to a differential equation exists and being able to write its solution in closed form.

Example 14.7 (Unsolvable ODE)**.** Consider the equation $y' = 2y/t$, with $y(0) \neq 0$ given; we are not dividing by zero because $y(0)$ is prescribed. Rewriting as

$$\frac{1}{y}\frac{dy}{dt} = \frac{2}{t}$$

and integrating with respect to t on both sides shows:

$$\ln|y| = 2\ln t + c$$

Or equivalently, $y = Ct^2$ for some $C \in \mathbb{R}$. In this expression, $y(0) = 0$, contradicting our initial conditions. Thus, this ODE has no solution with the given initial conditions.

Example 14.8 (Nonunique solutions)**.** Now, consider the same ODE with $y(0) = 0$. Consider $y(t)$ given by $y(t) = Ct^2$ for *any* $C \in \mathbb{R}$. Then, $y'(t) = 2Ct$. Thus,

$$\frac{2y}{t} = \frac{2Ct^2}{t} = 2Ct = y'(t),$$

showing that the ODE is solved by this function regardless of C. Thus, solutions of this equation with the new initial conditions are nonunique.

Thankfully, there is a rich theory characterizing behavior and stability of solutions to ordinary differential equations. Our development in the next chapter will have a stronger set of conditions needed for existence of a solution, but under weak conditions on f it is possible to show that an ODE $f'(t) = F[f(t)]$ has a solution. For instance, one such theorem guarantees local existence of a solution:

Theorem 14.1 (Local existence and uniqueness)**.** Suppose F is continuous and Lipschitz, that is, $\|F[\vec{y}] - F[\vec{x}]\|_2 \leq L\|\vec{y} - \vec{x}\|_2$ for some L. Then, the ODE $f'(t) = F[f(t)]$ admits exactly one solution for all $t \geq 0$ regardless of initial conditions.

In our subsequent development, we will assume that the ODE we are attempting to solve satisfies the conditions of such a theorem. This assumption is realistic since the conditions for existence and uniqueness are so weak.

14.2.3 Model Equations

One way to understand computational methods for integrating ODEs is to examine their behavior on well-understood *model equations*. Many ODEs locally can be approximated by these model equations, motivating our detailed examination of these simplistic test cases.

We start by introducing a model equation for ODEs with a single dependent variable. Given our simplifications in §14.2.1, we consider equations of the form $y' = F[y]$, where

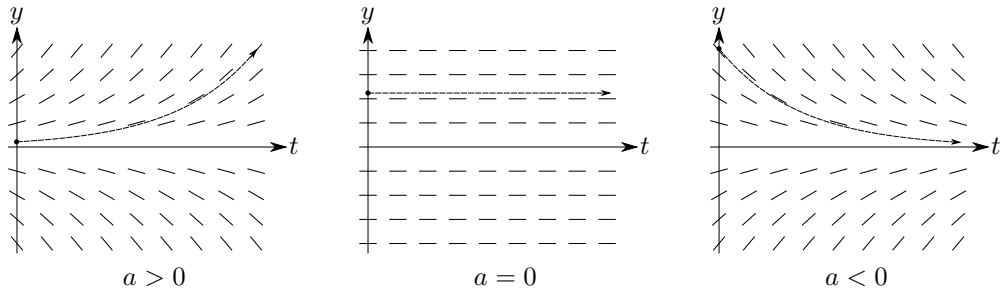


FIGURE 14.4 Three cases of the linear model equation $y' = ay$.

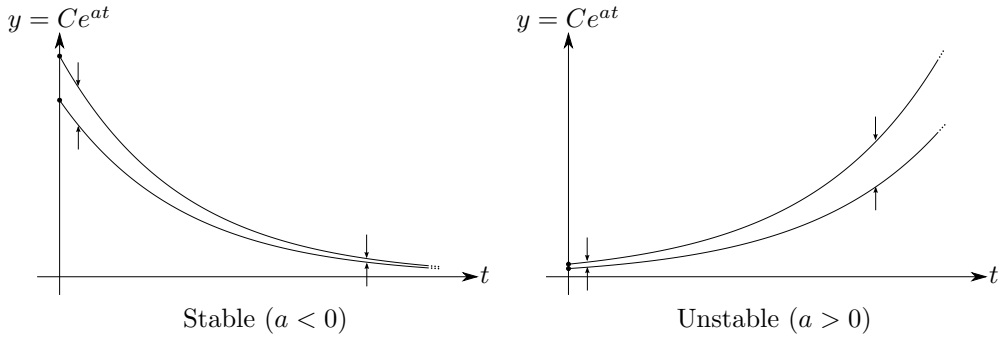


FIGURE 14.5 A stable ODE diminishes the difference between solutions over time t if $y(0)$ is perturbed, while an unstable ODE amplifies this difference.

$y(t) : [0, \infty) \to \mathbb{R}$. Taking a linear approximation of F, we might define $y' = ay + b$ to be the model equation, but we actually can fix $b = 0$. To justify using just one degree of freedom, define $\bar{y} \equiv y + b/a$. Then,

$$\begin{aligned}
\bar{y}' &= \left(y + \frac{b}{a}\right)' \text{ by definition of } \bar{y} \\
&= y' \text{ since the second term is constant with respect to } t \\
&= ay + b \text{ from the linearization} \\
&= a(\bar{y} - b/a) + b \text{ by inverting the definition of } \bar{y} \\
&= a\bar{y}.
\end{aligned}$$

Thus, $\bar{y}' = a\bar{y}$, showing that the constant b does not affect the qualitative behavior of the ODE. Hence, in the phenomenological study of model equations we safely take $b = 0$.

By the argument above, we locally can understand behavior of $y' = F[y]$ by studying the linear equation $y' = ay$. While the original ODE may not be solvable in closed form, applying standard arguments from calculus shows that the model equation is solved by the formula

$$y(t) = Ce^{at}.$$

Qualitatively, this formula splits into three cases, illustrated in Figure 14.4:

1. $a > 0$: In this case solutions get larger and larger; if $y(t)$ and $\hat{y}(t)$ both satisfy the ODE with slightly different starting conditions, as $t \to \infty$ they diverge.

2. $a = 0$: The system in this case is solved by constants; solutions with different starting points stay the same distance apart.

3. $a < 0$: Then, all solutions of the ODE approach 0 as $t \to \infty$.

We say cases 2 and 3 are *stable*, in the sense that perturbing $y(0)$ yields solutions that do not diverge from each other over time; case 1 is *unstable*, since a small mistake in specifying the initial condition $y(0)$ will be amplified as time t advances. Unstable ODEs generate ill-posed computational problems; without careful consideration we cannot expect numerical methods to generate usable solutions in this case, since even theoretical outputs are so sensitive to perturbations of the input. On the other hand, stable problems are well-posed since small mistakes in $y(0)$ get diminished over time. Both cases are shown in Figure 14.5.

Advancing to multiple dimensions, we could study the linearized equation $\vec{y}' = A\vec{y}$; for simplicity, we will assume A is symmetric. As explained in §5.1.2, if $\vec{y}_1, \cdots, \vec{y}_k$ are eigenvectors of A with eigenvalues $\lambda_1, \ldots, \lambda_k$ and $\vec{y}(0) = c_1\vec{y}_1 + \cdots + c_k\vec{y}_k$, then $\vec{y}(t) = c_1e^{\lambda_1 t}\vec{y}_1 + \cdots + c_ke^{\lambda_k t}\vec{y}_k$. Based on this formula, the eigenvalues of A take the place of a in the one-dimensional model equation. From this result, it is not hard to intuit that a multivariable $\vec{y}' = A\vec{y}$ is stable exactly when its spectral radius is at most one, that is, when all the eigenvalues of A have absolute value upper-bounded by one.

As in the single-variable case, in reality we likely wish to solve $\vec{y}' = F[\vec{y}]$ for general functions F. Assuming F is differentiable, we can write $F[\vec{y}] \approx F[\vec{y}_0] + J_F(\vec{y}_0)(\vec{y} - \vec{y}_0)$, yielding the model equation above after a shift; this shows that for short periods of time we expect behavior similar to the model equation with $A = J_F(\vec{y}_0)$, the Jacobian at $\vec{y}_0$.

14.3 TIME-STEPPING SCHEMES

We now describe several methods for solving the nonlinear ODE $\vec{y}' = F[\vec{y}]$ for potentially nonlinear functions F. Given a "time step" h, our methods will be used to generate estimates of $\vec{y}(t + h)$ given $\vec{y}(t)$ and F. Applying these methods iteratively generates estimates $\vec{y}_0 \equiv \vec{y}(t)$, $\vec{y}_1 \approx \vec{y}(t+h)$, $\vec{y}_2 \approx \vec{y}(t+2h)$, $\vec{y}_3 \approx \vec{y}(t+3h)$, and so on. We call methods for generating approximations of $\vec{y}(t)$ *time-stepping schemes* or *integrators*, reflecting the fact that they are integrating out the derivatives in the input equation.

Of key importance to our consideration is the idea of *stability*. Even if an ODE theoretically is stable using the definition from §14.2.3, the integrator may produce approximations that diverge at an exponential rate. Stability usually depends on the time step h; when h is too large, differential estimates of the quality of an integrator fail to hold, yielding unpredictable output. Stability, however, can compete with *accuracy*; stable schemes may generate bad approximations of $\vec{y}(t)$, even if they are guaranteed not to have wild behavior. ODE integrators that are both stable and accurate tend to require excessive computation time, indicating that we must compromise between these two properties.

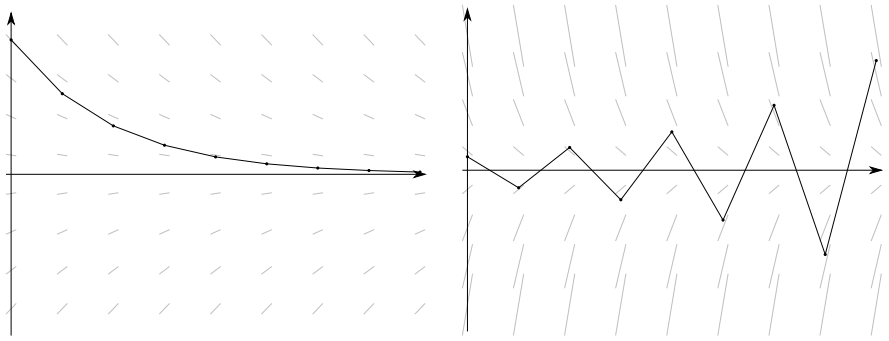
14.3.1 Forward Euler

Our first ODE integrator comes from our construction of the forward differencing scheme in §13.3.2:

$$F[\vec{y}_k] = \vec{y}'(t) = \frac{\vec{y}_{k+1} - \vec{y}_k}{h} + O(h)$$

Solving this relationship for $\vec{y}_{k+1}$ shows

$$\boxed{\vec{y}_{k+1} = \vec{y}_k + hF[\vec{y}_k] + O(h^2) \approx \vec{y}_k + hF[\vec{y}_k].}$$



Stable ($a = -0.4$) Unstable ($a = -2.3$)

FIGURE 14.6 Unstable and stable cases of forward Euler integration for the model equation $y' = ay$ with $h = 1$.

The *forward Euler* scheme applies the approximation on the right to estimate $\vec{y}_{k+1}$. It is one of the most computationally-efficient strategies for time-stepping.In particular, we call it an *explicit* method, that is, there is an explicit formula for $\vec{y}_{k+1}$ in terms of $\vec{y}_k$ and F.

Our approximation of $\vec{y}_{k+1}$ is $O(h^2)$, so each step induces quadratic error. We call this error *localized truncation error* because it is the error induced by a single step; the word "truncation" refers to the fact that we truncated a Taylor series to obtain this formula. Our iterate $\vec{y}_k$, however, already may be inaccurate thanks to accumulated truncation errors from previous iterations. If we integrate from t_0 to t with $O(1/h)$ steps, then our total error looks like $O(h)$; this estimate represents *global truncation error*, and thus we usually write that the forward Euler scheme is "first-order accurate."

The stability of forward Euler can be motivated by studying the model equation. In our discussion, we will work out the stability of methods in the one-variable case $y' = ay$, with the intuition that similar statements carry over to multidimensional equations by replacing a with a spectral radius. Substituting the one-variable model equation into the forward Euler scheme, we can write

$$y_{k+1} = y_k + ahy_k = (1 + ah)y_k.$$

Applying this formula recursively shows $y_k = (1 + ah)^k y_0$. Using this explicit formula for y_k in terms of y_0, we find that the integrator is stable when $|1 + ah| \leq 1$, since otherwise $|y_k| \to \infty$ exponentially. Assuming $a < 0$ (otherwise the problem is ill-posed), we can write this condition in a simpler form:

$$\begin{aligned} |1 + ah| \leq 1 &\iff -1 \leq 1 + ah \leq 1 \text{ by expanding the absolute value} \\ &\iff -2 \leq ah \leq 0 \\ &\iff 0 \leq h \leq \frac{2}{|a|}, \text{ since } a < 0. \end{aligned}$$

This derivation shows that forward Euler admits a *time step restriction* for stability. That is, the output of forward Euler can explode even when $y' = ay$ is stable when h is too large.

Figure 14.6 illustrates what happens when this condition is obeyed or violated. When time steps are too large—or equivalently when $|a|$ is too large—the forward Euler method is not only inaccurate but also has very different qualitative behavior. For nonlinear ODEs this formula gives a guide for stability at least locally in time; globally h may have to be adjusted if the Jacobian of F becomes worse conditioned.

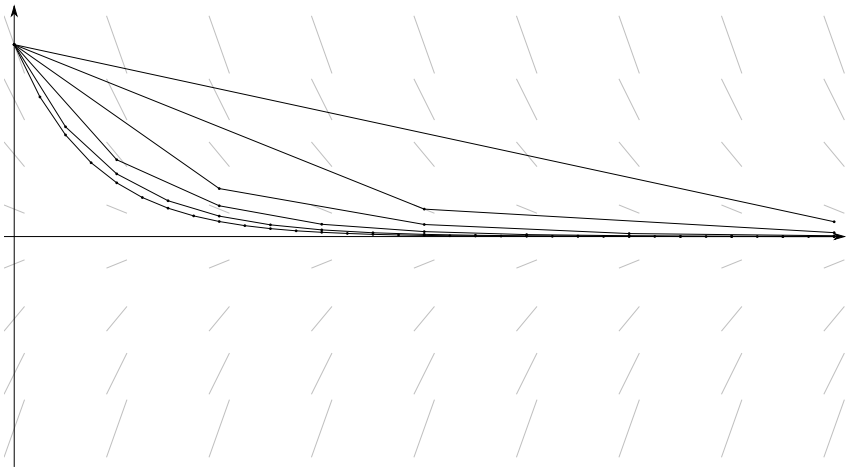


FIGURE 14.7 Backward Euler integration is unconditionally stable, so no matter how large a time step h with the same initial condition, the resulting approximate solution of $y' = ay$ does not diverge. While the output is stable, when h is large the result does not approximate the continuous solution $y = Ce^{at}$ effectively.

Certain well-posed ODEs require unreasonably small time steps h for forward Euler to be stable. In this case, even though the forward Euler formula is computationally inexpensive for a single step, integrating to some fixed time t may be infeasible because so many steps are needed. Such ODEs are called *stiff* equations, inspired by stiff springs that require tiny time steps to capture their rapid oscillations. One text defines stiff problems slightly differently (via [56]):

> "Stiff equations are problems for which explicit methods don't work." [53]

With this definition in mind, in the next section we consider an *implicit* method with no stability time step restriction, making it more suitable for stiff problems.

14.3.2 Backward Euler

Similarly, we could have applied the backward differencing scheme at $\vec{y}_{k+1}$ to design an ODE integrator:

$$F[\vec{y}_{k+1}] = \vec{y}'(t+h) = \frac{\vec{y}_{k+1} - \vec{y}_k}{h} + O(h)$$

Again isolating $\vec{y}_k$ shows that this integrator requires solving the following potentially nonlinear system of equations for $\vec{y}_{k+1}$:

$$\boxed{\vec{y}_{k+1} = \vec{y}_k + hF[\vec{y}_{k+1}].}$$

This equation differs from forward Euler integration by the evaluation of F at $\vec{y}_{k+1}$ rather than at $\vec{y}_k$. Because we have to solve this equation for $\vec{y}_{k+1}$, this technique, known as *backward Euler* integration, is an *implicit* integrator.

This method is first-order accurate like forward Euler by an identical argument. The

stability of this method, however, contrasts considerably with forward Euler. Once again considering the model equation $y' = ay$, we write:

$$y_k = y_{k+1} - hay_{k+1} \implies y_{k+1} = \frac{y_k}{1-ha}.$$

To prevent exponential blowup, we enforce the following condition:

$$\begin{aligned}\frac{1}{|1-ha|} \leq 1 &\iff |1-ha| \geq 1 \\ &\iff 1-ha \leq -1 \text{ or } 1-ha \geq 1 \\ &\iff h \leq \frac{2}{a} \text{ or } h \geq 0, \text{ for } a < 0\end{aligned}$$

We always take $h \geq 0$, so backward Euler is *unconditionally* stable, illustrated in Figure 14.7.

Even if backward Euler is stable, however, it may not be accurate. If h is too large, $\vec{y}_k$ will approach zero at the wrong rate. As a result, when simulating cloth and other physical materials that require lots of high-frequency detail to be realistic, backward Euler may exhibit undesirable dampening. Furthermore, we have to invert $F[\cdot]$ to solve for $\vec{y}_{k+1}$.

Example 14.9 (Backward Euler)**.** Suppose we wish to solve $\vec{y}' = A\vec{y}$ for $A \in \mathbb{R}^{n\times n}$. Then, to find $\vec{y}_{k+1}$ we solve the following system:

$$\vec{y}_k = \vec{y}_{k+1} - hA\vec{y}_{k+1} \implies \vec{y}_{k+1} = (I_{n\times n} - hA)^{-1}\vec{y}_k.$$

14.3.3 Trapezoidal Method

Suppose that in addition to having $\vec{y}_k$ at time t and $\vec{y}_{k+1}$ at time $t+h$, we also know $\vec{y}_{k+1/2}$ at the halfway point in time $t + h/2$. Then, by our derivation of the centered differencing we can write:

$$\vec{y}_{k+1} = \vec{y}_k + hF[\vec{y}_{k+1/2}] + O(h^3)$$

In our derivation of error bounds for the trapezoidal rule in §13.2.3, we derived the following relationship via Taylor's theorem:

$$\frac{F[\vec{y}_{k+1}] + F[\vec{y}_k]}{2} = F[\vec{y}_{k+1/2}] + O(h^2).$$

Substituting this equality into the expression for $\vec{y}_{k+1}$ yields our first second-order integration scheme, the *trapezoid method* for integrating ODEs:

$$\boxed{\vec{y}_{k+1} = \vec{y}_k + h\frac{F[\vec{y}_{k+1}] + F[\vec{y}_k]}{2}}$$

Like backward Euler, this method is implicit since we must solve this equation for $\vec{y}_{k+1}$.

To carry out stability analysis on $y' = ay$, we find in this case time steps of the trapezoidal method solve

$$y_{k+1} = y_k + \frac{1}{2}ha(y_{k+1} + y_k)$$

In other words,

$$y_k = \left(\frac{1 + \frac{1}{2}ha}{1 - \frac{1}{2}ha}\right)^k y_0.$$

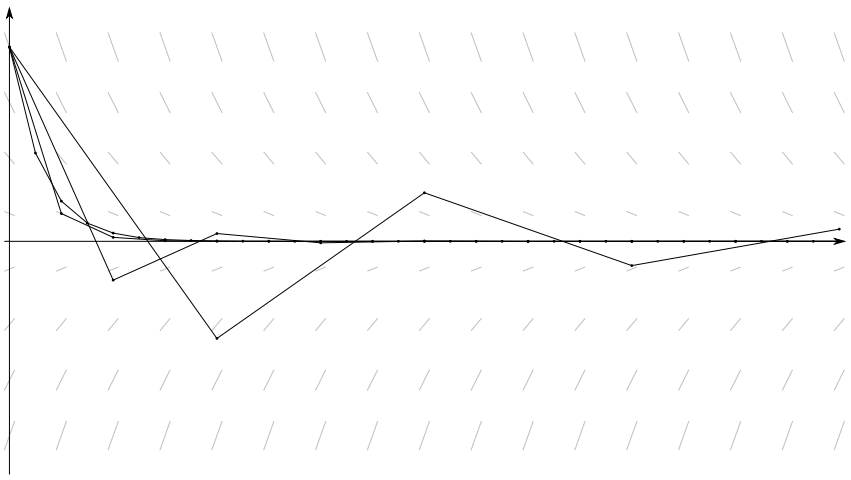


FIGURE 14.8 The trapezoidal method is unconditionally stable, so regardless of the step size h the solution curves always approach $y = 0$; when h is large, however, the output oscillates about zero as it decays.

The method is thus stable when

$$\left|\frac{1 + \frac{1}{2}ha}{1 - \frac{1}{2}ha}\right| < 1.$$

It is easy to see that this inequality holds whenever $a < 0$ and $h > 0$, showing that the trapezoid method is unconditionally stable.

Despite its higher order of accuracy with maintained stability, the trapezoid method has some drawbacks that make it less popular than backward Euler for large time steps. In particular, consider the ratio

$$R \equiv \frac{y_{k+1}}{y_k} = \frac{1 + \frac{1}{2}ha}{1 - \frac{1}{2}ha}.$$

When $a < 0$, for large enough h this ratio eventually becomes negative; as $h \to \infty$, we have $R \to -1$. Thus, as illustrated in Figure 14.8, if time steps are too large, the trapezoidal method of integration tends to introduce undesirable oscillatory behavior not present in theoretical solutions Ce^{at} of $y' = ay$.

14.3.4 Runge-Kutta Methods

A class of integrators can be derived by making the following observation:

$$\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + \int_{t_k}^{t_k+h} \vec{y}'(t)\,dt \text{ by the Fundamental Theorem of Calculus} \\ &= \vec{y}_k + \int_{t_k}^{t_k+h} F[\vec{y}(t)]\,dt \text{ since } \vec{y} \text{ satisfies } \vec{y}'(t) = F[\vec{y}(t)].\end{aligned}$$

Using this formula outright does not help design a method for time-stepping, since we do not know $\vec{y}(t)$ *a priori*. Approximating the integral using quadrature rules from the previous chapter, however, produces a class of well-known strategies for ODE integration.

For example, suppose we apply the trapezoidal quadrature rule to the integral for $\vec{y}_{k+1}$. This substitution shows:

$$\vec{y}_{k+1} = \vec{y}_k + \frac{h}{2}(F[\vec{y}_k] + F[\vec{y}_{k+1}]) + O(h^3)$$

This is the formula we wrote for the trapezoidal method in §14.3.3.

If we wish to find an explicit rather than implicit method with the accuracy of the trapezoidal time-stepping, however, we must replace $F[\vec{y}_{k+1}]$ with a high-accuracy approximation that is easier to evaluate:

$$\begin{aligned} F[\vec{y}_{k+1}] &= F[\vec{y}_k + hF[\vec{y}_k] + O(h^2)] \text{ by the forward Euler order of accuracy} \\ &= F[\vec{y}_k + hF[\vec{y}_k]] + O(h^2) \text{ by Taylor's theorem.} \end{aligned}$$

Since it gets scaled by h, making this substitution for $\vec{y}_{k+1}$ does not affect the order of approximation of the trapezoidal time step above. This change results in a new approximation:

$$\boxed{\vec{y}_{k+1} = \vec{y}_k + \frac{h}{2}(F[\vec{y}_k] + F[\vec{y}_k + hF[\vec{y}_k]]) + O(h^3).}$$

Ignoring the $O(h^3)$ terms yields a new integrator known as *Heun's method*, which is second-order accurate and explicit.

To study stability of Heun's method for the model equation $y' = ay$ with $a < 0$, we write:

$$y_{k+1} = y_k + \frac{h}{2}(ay_k + a(y_k + hay_k)) = \left(\frac{1}{2}h^2a^2 + ha + 1\right) y_k$$

Thus, the method is stable when

$$-1 \leq 1 + ha + \frac{1}{2}h^2a^2 \leq 1 \iff -4 \leq 2ha + h^2a^2 \leq 0$$

The inequality on the right shows $h \leq \frac{2}{|a|}$, and the inequality on the left is always true for $h > 0$ and $a < 0$. Hence, the stability condition for Heun's method can be written $h \leq \frac{2}{|a|}$, the same as the stability condition for Newton's method.

Heun's method is an example of a *Runge-Kutta* integrator derived by applying quadrature and substituting Euler steps for $F[\vec{y}_k + \ell h]$, for $\ell > 0$ as above. Forward Euler is a first-order accurate Runge-Kutta method, and Heun's method is second-order. A popular fourth-order Runge-Kutta method (abbreviated "RK4") is given by:

$$\boxed{\begin{aligned} \vec{y}_{k+1} &= \vec{y}_k + \frac{h}{6}(\vec{k}_1 + 2\vec{k}_2 + 2\vec{k}_3 + \vec{k}_4) \\ \text{where } \vec{k}_1 &= F[\vec{y}_k] \\ \vec{k}_2 &= F\left[\vec{y}_k + \frac{1}{2}h\vec{k}_1\right] \\ \vec{k}_3 &= F\left[\vec{y}_k + \frac{1}{2}h\vec{k}_2\right] \\ \vec{k}_4 &= F\left[\vec{y}_k + h\vec{k}_3\right] \end{aligned}}$$

This method arises from application of Simpson's quadrature rule.

Runge-Kutta methods are popular because they are explicit but provide high degrees of accuracy. The cost of this accuracy, however, is that $F[\cdot]$ must be evaluated more times. Furthermore, Runge-Kutta strategies can be extended to implicit methods when ODEs are poorly conditioned.

14.3.5 Exponential Integrators

We have focused our stability and accuracy analyses on the model equation $y' = ay$. If this ODE is truly an influential test case, however, we have neglected a key piece of information: We know the solution of $y' = ay$ in closed form as $y = Ce^{at}$! We might as well incorporate this formula into an integration scheme to achieve 100% accuracy on the model equation. That is, we can design a class of integrators that achieves strong accuracy when $F[\cdot]$ is nearly linear, potentially at the cost of computational efficiency.

Assuming A is symmetric, using the eigenvector method from §14.2.3 we can write the solution of the ODE $\vec{y}' = A\vec{y}$ as $\vec{y}(t) = e^{At}\vec{y}(0)$, where e^{At} is a matrix encoding the transformation from $\vec{y}(0)$ to $\vec{y}(t)$ (see problem 5.9). Starting from this formula, integrating in time by writing $\vec{y}_{k+1} = e^{Ah}\vec{y}_k$ achieves perfect accuracy on the linear model equation; our strategy is to apply this formula to the nonlinear case.

When F is smooth, we can attempt to factor the ODE $\vec{y}' = F[\vec{y}]$ as

$$\vec{y}' = A\vec{y} + G[\vec{y}],$$

where G is a nonlinear but small function; a typical approach is to use the Taylor expansion of F. Then, the class of *exponential integrators* integrate the $A\vec{y}$ part using the exponential formula and approximate the effect of the nonlinear G part separately.

We start by writing a "variation of parameters" formula. We rewrite the original ODE as $\vec{y}' - A\vec{y} = G[\vec{y}]$. Then, suppose we multiply both sides of this formula by e^{-At}. The resulting left hand side satisfies:

$$e^{-At}(\vec{y}' - A\vec{y}) = \frac{d}{dt}\left(e^{-At}\vec{y}(t)\right),$$

after applying the identity $Ae^{At} = e^{At}A$ (see problem 14.2). Hence, integrating from 0 to t shows

$$e^{-At}\vec{y}(t) - \vec{y}(0) = \int_0^t e^{-A\tau}G[\vec{y}(\tau)]\,d\tau$$

Or equivalently,

$$\begin{aligned}\vec{y}(t) &= e^{At}\vec{y}(0) + e^{At}\int_0^t e^{-A\tau}G[\vec{y}(\tau)]\,d\tau \\ &= e^{At}\vec{y}(0) + \int_0^t e^{A(t-\tau)}G[\vec{y}(\tau)]\,d\tau.\end{aligned}$$

Generalizing this formula slightly shows:

$$\vec{y}_{k+1} = e^{Ah}\vec{y}_k + \int_{t_k}^{t_k+h} e^{A(t_k+h-t)}G[\vec{y}(t)]\,dt$$

Similar to our derivation of the Runge-Kutta methods, exponential integrators apply quadrature to the integral on the right-hand side to approximate the time step to $\vec{y}_{k+1}$.

For example, the *first-order exponential integrator* applies forward Euler to the nonlinear G term by assuming the constant approximation $G[\vec{y}(t)] \approx G[\vec{y}_k]$:

$$\boxed{\vec{y}_{k+1} = e^{Ah}\vec{y}_k + A^{-1}(e^{Ah} - I_{n\times n})G[\vec{y}_k]}$$

Analyzing exponential integrators like this one requires techniques beyond using the linear model equation, since these integrators are designed to integrate linear ODEs exactly. Intuitively, they behave best when $G \approx 0$, but the cost of this high numerical performance is the use of the matrix exponential, which is difficult to apply efficiently.

14.4 MULTIVALUE METHODS

The transformations in §14.2.1 reduced all explicit ODEs to the form $\vec{y}' = F[\vec{y}]$, which can be integrated using the methods introduced in the previous section. While all explicit ODEs *can* be written this way, however, it is not clear that they always *should* be when designing a high-accuracy integrator.

When we reduced k-th order ODEs to first order, we introduced new variables representing the first through $(k-1)$-st derivatives of the desired output function $\vec{y}(t)$. The integrators in the previous section then approximate $\vec{y}(t)$ and these $k-1$ derivatives with equal accuracy, since in some sense they are treated "democratically" in first-order form. A natural question is whether we can relax the accuracy of the approximated derivatives of $\vec{y}(t)$ without affecting the quality of the $\vec{y}(t)$ estimate itself.

To support this perspective, consider the Taylor series

$$\vec{y}(t_k + h) = \vec{y}(t_k) + h\vec{y}'(t_k) + \frac{h^2}{2}\vec{y}''(t_k) + O(h^3)$$

If we perturb our estimate of $\vec{y}'(t_k)$ by some value on the order $O(h^2)$, the quality of the approximation does not change, since

$$h[\vec{y}'(t_k) + O(h^2)] = h\vec{y}'(t_k) + O(h^3).$$

Perturbing $\vec{y}''(t_k)$ by a value on the order $O(h)$ has a similar effect, since

$$\frac{h^2}{2}[\vec{y}''(t_k) + O(h)] = \frac{h^2}{2}\vec{y}''(t_k) + O(h^3).$$

Based on this Taylor series argument, *multivalue methods* integrate $\vec{y}^{(k)}(t) = F[t, \vec{y}'(t), \vec{y}''(t), \ldots, \vec{y}^{(k-1)}(t)]$ using less accurate estimates of the higher-order derivatives of $\vec{y}(t)$.

We will restrict our discussion to the second-order case $\vec{y}''(t) = F[t, \vec{y}, \vec{y}']$, the most common case for ODE integration thanks to Newton's second law $F = ma$. Extending the methods we consider to higher order, however, follows similar if notationally more complex arguments. For the remainder of this section, we will define a "velocity" vector $\vec{v}(t) \equiv \vec{y}'(t)$ and an "acceleration" vector $\vec{a} \equiv \vec{y}''(t)$. By the reduction to first order, we wish to solve the following order system:

$$\begin{aligned}\vec{y}'(t) &= \vec{v}(t)\\ \vec{v}'(t) &= \vec{a}(t)\\ \vec{a}(t) &= F[t, \vec{y}(t), \vec{v}(t)]\end{aligned}$$

Our goal is to derive integrators tailored to this system, evaluated based on the accuracy of estimating $\vec{y}(t)$ rather than $\vec{v}(t)$ or $\vec{a}(t)$.

14.4.1 Newmark Schemes

We begin by deriving the class of *Newmark* integrators following the development in [43]. Denote $\vec{y}_k$, $\vec{v}_k$, and $\vec{a}_k$ as position, velocity, and acceleration vectors at time t_k; our goal is to advance to time $t_{k+1} \equiv t_k + h$. By the Fundamental Theorem of Calculus we can write

$$\vec{v}_{k+1} = \vec{v}_k + \int_{t_k}^{t_{k+1}} \vec{a}(t)\, dt$$

We also can write $\vec{y}_{k+1}$ as an integral involving $\vec{a}(t)$, by writing the same error estimate developed in some proofs of Taylor's theorem:

$$\begin{aligned}
\vec{y}_{k+1} &= \vec{y}_k + \int_{t_k}^{t_{k+1}} \vec{v}(t)\,dt \text{ by the Fundamental Theorem of Calculus} \\
&= \vec{y}_k + [t\vec{v}(t)]_{t_k}^{t_{k+1}} - \int_{t_k}^{t_{k+1}} t\vec{a}(t)\,dt \text{ after integration by parts} \\
&= \vec{y}_k + t_{k+1}\vec{v}_{k+1} - t_k\vec{v}_k - \int_{t_k}^{t_{k+1}} t\vec{a}(t)\,dt \text{ by expanding the difference term} \\
&= \vec{y}_k + h\vec{v}_k + t_{k+1}\vec{v}_{k+1} - t_{k+1}\vec{v}_k - \int_{t_k}^{t_{k+1}} t\vec{a}(t)\,dt \text{ by adding and subtracting } h\vec{v}_k \\
&= \vec{y}_k + h\vec{v}_k + t_{k+1}(\vec{v}_{k+1} - \vec{v}_k) - \int_{t_k}^{t_{k+1}} t\vec{a}(t)\,dt \text{ after factoring} \\
&= \vec{y}_k + h\vec{v}_k + t_{k+1}\int_{t_k}^{t_{k+1}} \vec{a}(t)\,dt - \int_{t_k}^{t_{k+1}} t\vec{a}(t)\,dt \text{ since } \vec{v}'(t) = \vec{a}(t) \\
&= \vec{y}_k + h\vec{v}_k + \int_{t_k}^{t_{k+1}} (t_{k+1} - t)\vec{a}(t)\,dt
\end{aligned}$$

Suppose we choose $\tau \in [t_k, t_{k+1}]$. Then, we can write expressions for $\vec{a}_k$ and $\vec{a}_{k+1}$ using the Taylor series about τ:

$$\begin{aligned}
\vec{a}_k &= \vec{a}(\tau) + \vec{a}'(\tau)(t_k - \tau) + O(h^2) \\
\vec{a}_{k+1} &= \vec{a}(\tau) + \vec{a}'(\tau)(t_{k+1} - \tau) + O(h^2)
\end{aligned}$$

For any constant $\gamma \in \mathbb{R}$, scaling the expression for $\vec{a}_k$ by $1 - \gamma$, scaling the expression for $\vec{a}_{k+1}$ by γ, and summing shows

$$\begin{aligned}
\vec{a}(\tau) &= (1-\gamma)\vec{a}_k + \gamma\vec{a}_{k+1} + \vec{a}'(\tau)((\gamma - 1)(t_k - \tau) - \gamma(t_{k+1} - \tau)) + O(h^2) \\
&= (1-\gamma)\vec{a}_k + \gamma\vec{a}_{k+1} + \vec{a}'(\tau)(\tau - h\gamma - t_k) + O(h^2) \text{ after substituting } t_{k+1} = t_k + h.
\end{aligned}$$

To integrate $\vec{a}(t)$ from t_k to t_{k+1} to get the change in velocity, we can apply our new approximation:

$$\begin{aligned}
\int_{t_k}^{t_{k+1}} \vec{a}(\tau)\,d\tau &= (1-\gamma)h\vec{a}_k + \gamma h\vec{a}_{k+1} + \int_{t_k}^{t_{k+1}} \vec{a}'(\tau)(\tau - h\gamma - t_k)\,d\tau + O(h^3) \\
&= (1-\gamma)h\vec{a}_k + \gamma h\vec{a}_{k+1} + O(h^2),
\end{aligned}$$

where the second step holds because $(\tau - t_k) - h\gamma = O(h)$ for $\tau \in [t_k, t_{k+1}]$ and the interval of integration is of width h. Applying this formula, we know

$$\vec{v}_{k+1} = \vec{v}_k + (1-\gamma)h\vec{a}_k + \gamma h\vec{a}_{k+1} + O(h^2).$$

Starting again from the approximation we wrote for $\vec{a}(\tau)$—this time using a new constant β rather than γ—we can also develop an approximation for $\vec{y}_{k+1}$. To do so, we will work with the integrand in the Taylor estimate for $\vec{y}_{k+1}$:

$$\begin{aligned}
\int_{t_k}^{t_{k+1}} (t_{k+1} - t)\vec{a}(t)\,dt &= \int_{t_k}^{t_{k+1}} (t_{k+1} - \tau)((1-\beta)\vec{a}_k + \beta\vec{a}_{k+1} + \vec{a}'(\tau)(\tau - h\beta - t_k))\,d\tau + O(h^3) \\
&= \frac{1}{2}(1-\beta)h^2\vec{a}_k + \frac{1}{2}\beta h^2\vec{a}_{k+1} + O(h^2) \text{ by a similar simplification.}
\end{aligned}$$

Thus, our earlier relationship shows:

$$\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + h\vec{v}_k + \int_{t_k}^{t_{k+1}} (t_{k+1} - t)\vec{a}(t)\,dt \text{ from before} \\ &= \vec{y}_k + h\vec{v}_k + \left(\frac{1}{2} - \beta\right) h^2\vec{a}_k + \beta h^2\vec{a}_{k+1} + O(h^2)\end{aligned}$$

Summarizing this technical argument, we have derived the class of Newmark schemes, each characterized by the two fixed parameters γ and β:

$$\boxed{\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + h\vec{v}_k + \left(\frac{1}{2} - \beta\right) h^2\vec{a}_k + \beta h^2\vec{a}_{k+1} \\ \vec{v}_{k+1} &= \vec{v}_k + (1-\gamma)h\vec{a}_k + \gamma h\vec{a}_{k+1} \\ \vec{a}_k &= F[t_k, \vec{y}_k, \vec{v}_k]\end{aligned}}$$

This integrator is accurate up to $O(h^2)$ in each time step, making it globally first-order accurate. Depending on γ and β, the integrator can be implicit, since $\vec{a}_{k+1}$ appears in the expressions for $\vec{y}_{k+1}$ and $\vec{v}_{k+1}$.

Specific choices of β and γ yield integrators with additional properties:

- $\beta = \gamma = 0$ gives the *constant acceleration* integrator:

$$\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + h\vec{v}_k + \frac{1}{2}h^2\vec{a}_k \\ \vec{v}_{k+1} &= \vec{v}_k + h\vec{a}_k\end{aligned}$$

  This integrator is explicit and holds exactly when the acceleration is a constant function of time.

- $\beta = 1/2, \gamma = 1$ gives the *constant implicit acceleration* integrator:

$$\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + h\vec{v}_k + \frac{1}{2}h^2\vec{a}_{k+1} \\ \vec{v}_{k+1} &= \vec{v}_k + h\vec{a}_{k+1}\end{aligned}$$

  The velocity is stepped implicitly using backward Euler, giving first-order accuracy. The $\vec{y}$ update, however, can be written

$$\vec{y}_{k+1} = \vec{y}_k + \frac{1}{2}h(\vec{v}_k + \vec{v}_{k+1}),$$

  which coincides with the trapezoidal rule. Hence, this is our first example of a scheme where the velocity and position updates have different orders of accuracy. This technique, however, is still only globally first-order accurate in $\vec{y}$.

- $\beta = 1/4, \gamma = 1/2$ gives the following second-order trapezoidal scheme after some algebra:

$$\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + \frac{1}{2}h(\vec{v}_k + \vec{v}_{k+1}) \\ \vec{v}_{k+1} &= \vec{v}_k + \frac{1}{2}h(\vec{a}_k + \vec{a}_{k+1})\end{aligned}$$

- $\beta = 0, \gamma = 1/2$ gives a second-order accurate *central differencing* scheme. In the canonical form, it is written

$$\vec{y}_{k+1} = \vec{y}_k + h\vec{v}_k + \frac{1}{2}h^2\vec{a}_k$$
$$\vec{v}_{k+1} = \vec{v}_k + \frac{1}{2}h(\vec{a}_k + \vec{a}_{k+1}).$$

  The method earns its name because simplifying the equations above leads to the alternative form:

$$\vec{v}_{k+1} = \frac{\vec{y}_{k+2} - \vec{y}_k}{2h}$$
$$\vec{a}_{k+1} = \frac{\vec{y}_{k+2} - 2\vec{y}_{k+1} + \vec{y}_k}{h^2}$$

- Newmark integrators are unconditionally stable when $4\beta > 2\gamma > 1$, with second-order accuracy exactly when $\gamma = 1/2$.

14.4.2 Staggered Grid and Leapfrog

A different way to achieve second-order accuracy in stepping $\vec{y}$ is to use centered differences about $t_{k+1/2} \equiv t_k + h/2$:

$$\vec{y}_{k+1} = \vec{y}_k + h\vec{v}_{k+1/2}$$

Rather than attempting to approximate $\vec{v}_{k+1/2}$ from $\vec{v}_k$ and/or $\vec{v}_{k+1}$, we can process velocities $\vec{v}$ directly at *half* points on the grid of time steps.

A similar update steps forward the velocities with the same accuracy:

$$\vec{v}_{k+3/2} = \vec{v}_{k+1/2} + h\vec{a}_{k+1}.$$

A lower-order approximation suffices for the acceleration term since it is a higher-order derivative:

$$\vec{a}_{k+1} = F\left[t_{k+1}, \vec{x}_{k+1}, \frac{1}{2}(\vec{v}_{k+1/2} + \vec{v}_{k+3/2})\right]$$

This expression can be substituted into the equation for $\vec{v}_{k+3/2}$.

When $F[\cdot]$ has no dependence on $\vec{v}$, e.g. when simulating particles without wind resistance, the method is fully explicit:

$$\boxed{\begin{aligned}\vec{y}_{k+1} &= \vec{y}_k + h\vec{v}_{k+1/2}\\ \vec{a}_{k+1} &= F[t_{k+1}, \vec{y}_{k+1}]\\ \vec{v}_{k+3/2} &= \vec{v}_{k+1/2} + h\vec{a}_{k+1}\end{aligned}}$$

This is known as the *leapfrog* method of integration, thanks to the staggered grid of times and the fact that each midpoint is used to update the next velocity or position.

One distinguishing property of the leapfrog scheme is its *time reversibility*.* Assume we have used the leapfrog integrator to generate $(\vec{y}_{k+1}, \vec{v}_{k+3/2}, \vec{a}_{k+1})$. Starting at t_{k+1}, we might reverse the direction of time and try to step backward. The leapfrog equations give

$$\vec{v}_{k+1/2} = \vec{v}_{k+3/2} + (-h)\vec{a}_{k+1}$$
$$\vec{y}_k = \vec{y}_{k+1} - h\vec{v}_{k+1/2}.$$

*Discussion of time reversibility contributed by Julian Kates-Harbeck.

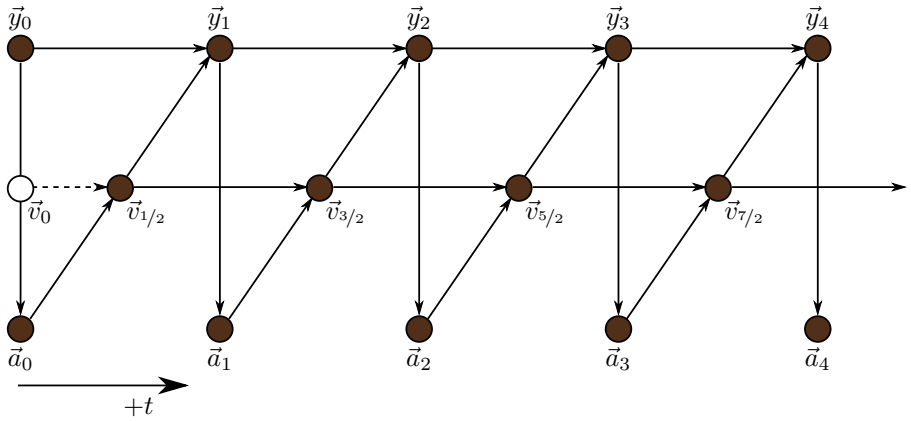


FIGURE 14.9 Explicit leapfrog integration computes velocities at half time steps; here arrows denote dependencies between the different computed values. If the initial conditions specify $\vec{v}$ at $t = 0$, an initial half time step must be carried out to approximate $\vec{v}_{1/2}$.

These formulas invert the forward time step equations. That is, if we run the leapfrog in reverse, we will trace our solution back to where we started *exactly* up to rounding error. This property comes from the symmetric form of the leapfrog scheme.

A consequence of reversibility is that errors in position, energy, and angular momentum tend to cancel out over time as opposed to accumulating. For instance, for problems where the acceleration only depends on position, angular momentum is conserved exactly by leapfrog integration, and energy stable over time, whereas other even higher-order schemes can induce significant "drift" of these quantities over time. Symmetry, second order accuracy for "first order work" (i.e. the same amount of computation as for Euler integration), and the conservation properties make the leapfrog integration a popular method for physical simulation.

If $F[\cdot]$ has dependence on $\vec{v}$, then this "staggered grid" method becomes implicit. Such dependence on velocity often is symmetric. For instance, wind resistance changes sign if you reverse the direction in which you are moving. This property makes the matrices symmetric in the implicit step for updating velocities, making it possible to use conjugate gradients and related fast iterative methods.

14.5 COMPARISON OF INTEGRATORS

This chapter has introduced a sampling from the remarkably large pantheon of ODE integrators. Choosing the right ODE for a given problem is a challenging task representing a careful balancing act between accuracy, stability, computational efficiency, and assorted special properties like reversibility. The table in Figure 14.10 compares the basic properties of these methods.

In practice, it may require some experimentation to determine the proper integrator given an ODE problem; thankfully, most of the integrators we have introduced are relatively easy to implement. In addition to the generic considerations we have discussed in this chapter, additional "domain-specific" concerns also influence the choice of ODE integrators, including the following:

Integrator	*Section*	*Accuracy*	*Implicit or explicit?*	*Stability*	*Notes*
Forward Euler	§14.3.1	First	Explicit	Conditional	
Backward Euler	§14.3.2	First	Implicit	Unconditional	
Trapezoidal	§14.3.3	Second	Implicit	Unconditional	Large steps oscillate
Heun	§14.3.4	Second	Explicit	Conditional	
RK4	§14.3.4	Fourth	Explicit	Conditional	
First-order exponential	§14.3.5	First	Explicit	Conditional	Requires matrix exponential
Newmark	§14.4.1	First	Implicit	Conditional	For 2nd-order ODE; 2nd-order accurate when $\gamma = 1/2$; explicit when $\beta = \gamma = 0$
Staggered	§14.4.2	Second	Implicit	Conditional	For 2nd-order ODE
Leapfrog	§14.4.2	Second	Explicit	Conditional	For 2nd-order ODE; reversible; $F[\cdot]$ must not depend on $\vec{v}$

FIGURE 14.10 Comparison of ODE integrators.

- In computer graphics and other fields prioritizing visual effect over reproducibility in the real world, it may be more important that a time-stepping method *looks* right than whether the numerical output is perfect. For instance, simulation tools for visual effects need to produce fluids, gases, and cloth that exhibit high-frequency swirls, vortices, and folds. These features may be dampened by a backward Euler integrator, even if it is more likely to be stable than other alternatives.
- Most of our analysis used Taylor series and other localized arguments, but *long-term behavior* of certain integrators can be favorable even if individual time steps are suboptimal. For instance, forward Euler integration tends to add energy to oscillatory ODEs, while backward Euler removes it. If we wish to simulate a pendulum swinging in perpetuity, neither of these techniques will suffice.
- Some ODEs operate in the presence of constraints. For instance, if we simulate a ball attached to a string, we may not wish for the string to stretch beyond its normal length. Methods like forward Euler and leapfrog integration can overshoot such constraints, so an additional projection step may be needed to enforce the constraints more exactly.
- A degree of adaptivity is needed for applications in which discrete events can happen during the course of solving an ODE. For instance, when simulating the dynamics of a piece of cloth, typically pieces of the cloth can run into each other or objects in their surroundings. These collision events can occur at fractional time steps and must be handled separately to avoid interpenetration of objects in a scene [5].
- For higher-quality animation and physical predictions, some ODE integrators can output not only the configuration at discrete time steps but also some indicator (e.g. an interpolatory formula) approximating continuous behavior between the time steps.

- If the function F in $\vec{y}' = F[\vec{y}]$ is smooth and differentiable, the derivatives of F can be used to improve the quality of time-stepping methods.

Many of these problems are difficult to handle efficiently in large-scale simulations and in other use cases where computational power is relatively limited.

14.6 EXERCISES

14.1 Some practice discretizing an ODE:

(a) Suppose we wish to solve the ODE $dy/dt = -\sin y$ numerically. For time step $h > 0$, write the implicit backward Euler equation for approximating y_{k+1} at $t = (k+1)h$ given y_k at $t = kh$.

(b) Write the Newton iteration for solving the equation from 14.1a for y_{k+1}.

14.2 We continue our discussion of the matrix exponential introduced in problem 5.9 and used in our discussion of exponential integrators. For this problem, assume $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix.

(a) Show that A commutes with e^{At} for any $t \geq 0$. That is, justify the formula $Ae^{At} = e^{At}A$.

(b) Recall that we can write

$$e^{At} = I_{n\times n} + At + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \cdots.$$

For sufficiently small $h \geq 0$ prove a similar formula for matrix inverses:

$$(I_{n\times n} - hA)^{-1} = I_{n\times n} + hA + (hA)^2 + (hA)^3 + \cdots$$

(c) Which of the two series from part 14.2b should converge faster? Based on this observation, compare the computational cost of a single backward Euler iteration (see Example 14.9) versus that of an iteration of the exponential integrator from §14.3.5 using these formulas.

14.3 Suppose we are solving a second-order ODE using the leapfrog integrator. We are given initial conditions $\vec{y}(0)$ and $\vec{v}(0)$, the position and velocity vectors at time $t = 0$. But, the leapfrog scheme maintains velocities at the half time steps. Propose a way to initialize $\vec{v}_{1/2}$ at time $t = h/2$, and argue that your initialization does not affect the order of accuracy of the leapfrog integrator if it is run for enough time steps.

14.4 Add together Taylor expansions for $\vec{y}(t+h)$ and $\vec{y}(t-h)$ to derive the *Verlet algorithm* for predicting $\vec{y}_{k+1}$ from $\vec{y}_k$ and $\vec{y}_{k-1}$. Show that this algorithm is equivalent to leapfrog integration and that a single step approximates $\vec{y}_{k+1}$ up to $O(h^4)$.

14.5 In this problem, we will motivate an ODE used in computer graphics applications that does not come from Newton's laws. Throughout this problem, assume $f, g : [0, 1] \to \mathbb{R}$ are differentiable functions with $g(0) = g(1) = 0$. We will derive continuous and discrete versions of the *screened Poisson equation*, used for smoothing (see e.g. [22]).

(a) So far our optimization problems have been to find points $\vec{x}^* \in \mathbb{R}^n$ minimizing some function $h(\vec{x})$, but sometimes our unknown is an entire function. Thankfully, our "variational" approach still is valid in this case. Explain in words what the following energies, which take a function f as input, measure about f:

(i) $E_1[f] \equiv \int_0^1 (f(t) - f_0(t))^2\,dt$ for some fixed function $f_0 : [0,1] \to \mathbb{R}$

(ii) $E_2[f] \equiv \int_0^1 (f'(t))^2\,dt$

(b) For an energy functional $E[\cdot]$ like the two above, explain how the following expression for $dE(f;g)$ (the Gâteaux derivative of E) can be thought of as the "directional derivative of E at f in the g direction:"

$$dE(f;g) = \frac{d}{d\varepsilon} E[f + \varepsilon g]\Big|_{\varepsilon=0}$$

(c) Again assuming $g(0) = g(1) = 0$, derive the following formulae:

(i) $dE_1(f,g) = \int_0^1 2(f(t) - f_0(t))g(t)\,dt$

(ii) $dE_2(f,g) = \int_0^1 -2f''(t)g(t)\,dt$
Hint: Apply integration by parts to get rid of $g'(t)$; recall our assumption $g(0) = g(1) = 0$.

(d) Suppose we wish to approximate f_0 with a smoother function f. One reasonable model for doing so is to minimize $E[f] \equiv E_1[f] + \alpha E_2[f]$ for some $\alpha > 0$ controlling the trade-off between similarity to f_0 and smoothness. Using the result of 14.5c, argue informally that an f minimizing this energy should satisfy the differential equation $f(t) - f_0(t) = \alpha f''(t)$ for $t \in (0,1)$.

(e) Now, suppose we discretize f on $[0,1]$ using n evenly-spaced samples $f^1, f^2, \ldots, f^n \in \mathbb{R}$ and f_0 using samples $f_0^1, f_0^2, \ldots, f_0^n$. Devise a discrete analog of $E[f]$ as a quadratic energy in the f^k's. For $k \notin \{1, n\}$, does differentiating E with respect to f_k yield a result analogous to 14.5d?

14.6 (adapted from [19]) The swing angle θ of a pendulum under gravity satisfies the following ODE:

$$\theta'' = -\sin\theta,$$

where $|\theta(0)| < \pi$ and $\theta'(0) = 0$.

(a) Suppose $\theta(t)$ solves the ODE. Show that the following value (representing the energy of the system) is constant as a function of t:

$$E(t) \equiv \frac{1}{2}(\theta')^2 - \cos\theta$$

(b) Many ODE integrators drift away from the desired output as time progresses over larger periods. For instance, forward Euler can add energy to a system by overshooting, while backward Euler tends to damp out motion and remove energy. In many computer graphics applications, quality long-term behavior can be prioritized, since large scale issues cause visual artifacts. The class of *symplectic* integrators is designed to avoid this issue.
Denote $\omega \equiv \theta'$. The *symplectic Euler* scheme makes a series of estimates

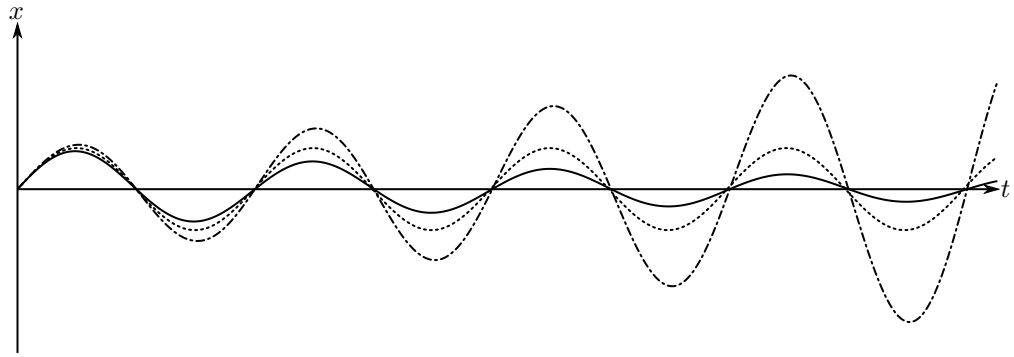


FIGURE 14.11 Three simulations of an undamped oscillator.

$\theta_0, \theta_1, \theta_2, \theta_3, \ldots$ and $\omega_0, \omega_1, \omega_2, \omega_3, \ldots$ at time $t = 0, h, 2h, 3h, \ldots$ using the following iteration:

$$\theta_{k+1} = \theta_k + h\omega_k$$
$$\omega_{k+1} = \omega_k - h \sin \theta_{k+1}.$$

Define

$$E_k \equiv \frac{1}{2}\omega_k^2 - \cos\theta_k.$$

Show that $E_{k+1} = E_k + O(h^2)$.

(c) Suppose we make the small-angle approximation $\sin\theta \approx \theta$ and decide to solve the linear ODE $\theta'' = -\theta$ instead. Now, symplectic Euler takes the following form:

$$\theta_{k+1} = \theta_k + h\omega_k$$
$$\omega_{k+1} = \omega_k - h\theta_{k+1}.$$

Write a 2×2 matrix A such that

$$\begin{pmatrix} \theta_{k+1} \\ \omega_{k+1} \end{pmatrix} = A \begin{pmatrix} \theta_k \\ \omega_k \end{pmatrix}.$$

(d) If we define $E_k \equiv \omega_k^2 + h\omega_k\theta_k + \theta_k^2$, show that $E_{k+1} = E_k$ in the iteration from 14.6c. In other words, E_k is constant from time step to time step.

14.7 Suppose we simulate a spring by solving the ODE $y'' = -y$ with $y(0) = 0$ and $y'(0) = 1$. We obtain the three plots of $y(t)$ in Figure 14.11 by using forward Euler, backward Euler, and symplectic Euler time integration. Determine which plot is which, and justify your answers using properties of the three integrators.

14.8 Suppose we discretize Schrödinger's equation for a particular quantum simulation yielding an ODE $\vec{x}' = A\vec{x}$, for $\vec{x}(t) \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$. Furthermore, suppose that A is *self-adjoint* and *negative definite*, that is, A satisfies the following properties:

- Self-adjoint: $a_{ij} = \bar{a}_{ji}$, where $\overline{a + bi} = a - bi$.

- Negative definite: $\vec{\bar{x}}^\top A\vec{x} \leq 0$ (and is real) for all $\vec{x} \in \mathbb{C}^n \backslash \{\vec{0}\}$. Here we define $(\vec{x})_i \equiv \bar{x}_i$.

Derive a backward Euler formula for solving this ODE and show that each step can be carried out using conjugate gradients.
Hint: Before discretizing, convert the ODE to a real-valued system by separating imaginary and real parts of the variables and constants.

14.9 ("Phi functions," [84]) Exponential integrators made use of ODEs with known solutions to boost numerical quality of time integration. This strategy can be extended using additional closed-form solutions.

(a) Define $\varphi_k(x)$ recursively by defining $\varphi_0(x) \equiv e^x$ and recursively writing

$$\varphi_{k+1}(x) \equiv \frac{1}{x}\left(\varphi_k(x) - \frac{1}{k!}\right).$$

Write the Taylor expansions of $\varphi_0(x)$, $\varphi_1(x)$, $\varphi_2(x)$, and $\varphi_3(x)$ about $x = 0$.

(b) Show that for $k \geq 1$,

$$\varphi_k(x) = \frac{1}{(k-1)!}\int_0^1 e^{(1-\theta)x}\theta^{k-1}\,d\theta.$$

Hint: Use integration by parts to show that the recursive relationship from 14.9a holds.

(c) Check the following formula for $\varphi_k'(x)$ when $k \geq 1$:

$$\varphi_k'(x) = \frac{1}{x}\left[\varphi_k(x)(x-k) + \frac{1}{(k-1)!}\right]$$

(d) Show that the ODE

$$\vec{u}'(t) = L\vec{u}(t) + \frac{t^k}{k!}\vec{u}_k$$

subject to $\vec{u}(0) = \vec{u}_0$ is solved by

$$\vec{u}(t) = \varphi_0(tL)\vec{u}_0 + \sum_{\ell=0}^{k}\frac{t^{\ell+1}}{(k-\ell)!}\varphi_{\ell+1}(tL)\vec{u}_k.$$

(e) Use this new closed-form solution to propose an exponential-type integrator for the ODE $\vec{y}' = A\vec{y} + \frac{t^k}{k!}\vec{u}_k + G[\vec{y}]$.

14.10 ("Fehlberg's method," [36] via notes by J. Feldman) We can approximate the error of an ODE integrator to help choose appropriate step sizes given a desired level of accuracy.

(a) Suppose we carry out a single time step of $\vec{y}' = F[\vec{y}]$ with size h starting from $\vec{y}(0) = \vec{y}_0$. Make the following definitions:

$$\begin{aligned}\vec{v}_1 &\equiv F[\vec{y}_0]\\ \vec{v}_2 &\equiv F[\vec{y}_0 + h\vec{v}_1]\\ \vec{v}_3 &\equiv F\left[\vec{y}_0 + \frac{h}{4}(\vec{v}_1 + \vec{v}_2)\right].\end{aligned}$$

We can write two estimates of $\vec{y}(h)$:

$$\vec{y}^{(1)} \equiv \vec{y}_0 + \frac{h}{2}(\vec{v}_1 + \vec{v}_2)$$
$$\vec{y}^{(2)} \equiv \vec{y}_0 + \frac{h}{6}(\vec{v}_1 + \vec{v}_2 + 4\vec{v}_3).$$

Show that there is some $K \in \mathbb{R}$ such that $\vec{y}^{(1)} = \vec{y}(h) + Kh^3 + O(h^4)$ and $\vec{y}^{(2)} = \vec{y}(h) + O(h^4)$.

(b) Use this relationship to derive an approximation of the amount of error introduced per unit increase of time t if we use $\vec{y}^{(1)}$ as an integrator. If this value is too large, adaptive integrators reject the step and try again with a smaller h.

CHAPTER 15

Partial Differential Equations

CONTENTS

15.1 Motivation 320
15.2 Statement and Structure of PDEs 325
15.2.1 Properties of PDEs 325
15.2.2 Boundary Conditions 326
15.3 Model Equations 328
15.3.1 Elliptic PDEs 328
15.3.2 Parabolic PDEs 329
15.3.3 Hyperbolic PDEs 330
15.4 Representing Derivative Operators 331
15.4.1 Finite Differences 332
15.4.2 Collocation 336
15.4.3 Finite Elements 337
15.4.4 Finite Volumes 340
15.4.5 Other Methods 341
15.5 Solving Parabolic and Hyperbolic Equations 342
15.5.1 Semidiscrete Methods 342
15.5.2 Fully Discrete Methods 343
15.6 Numerical Considerations 344
15.6.1 Consistency, Convergence, and Stability 344
15.6.2 Linear Solvers for PDE 345

INTUITION motivating much of the study of ordinary differential equations stems from the time evolution of physical systems. Equations like Newton's second law determining the motion of physical objects over time dominate the literature on ODE problems; additional examples come from chemical concentrations reacting over time, populations of predators and prey interacting from season to season, and so on. In each case, the initial configuration—e.g. the positions and velocities of particles in a system at time zero—is known, and the task is to predict behavior as time progresses. Derivatives only appear in a single time variable.

In this chapter, however, we entertain the possibility of *coupling* relationships between different derivatives of a function. It is not difficult to find examples where this coupling is necessary. When simulating gases or fluids, quantities like "pressure gradients," which encode the derivatives of pressure in *space*, figure into how material moves over *time*; these gradients appear since gases and fluids naturally move from high-pressure regions to low-pressure regions. In image processing, coupling the horizontal and vertical partial derivatives of an image can be used to describe its edges, characterize its texture, and so on.

Equations coupling together derivatives of functions are known as *partial differential*

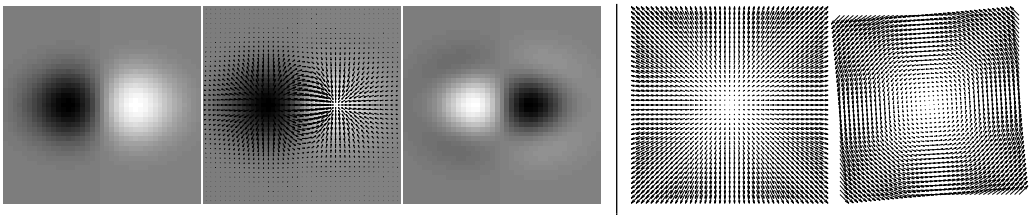


FIGURE 15.1 Vector calculus notation. On the left, we show a function $f(\vec{x})$ for $\vec{x} \in \mathbb{R}^2$ colored from black to white, its gradient ∇f, and its Laplacian $\nabla^2 f$; on the right are vector fields $\vec{v}(\vec{x})$ with different balances between divergence and curl.

equations. They are the subject of a rich, nuanced theory worthy of larger-scale treatment, so we simply will summarize key ideas and provide sufficient material to approach problems commonly appearing in practice.

15.1 MOTIVATION

Partial differential equations (PDEs) provide one or more relationships between the partial derivatives of a function $f : \mathbb{R}^n \to \mathbb{R}^m$; the goal is to find an f satisfying the criteria. PDEs appear in nearly any branch of applied mathematics, and we list just a few below. Unlike in previous chapters, the algorithms in this chapter will be far from optimal when applied to many of the examples. Rather, given the space available, our goals are to explore the vast space of problems that can be expressed as PDEs, to introduce the language needed to determine necessary numerical machinery, and to highlight key challenges and techniques for different classes of PDEs.

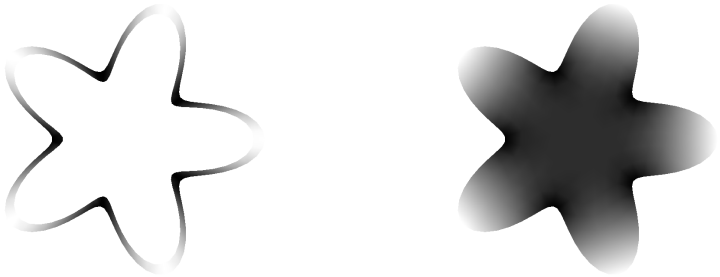
There are a few combinations of partial derivatives that appear often in the world of PDEs. If $f : \mathbb{R}^3 \to \mathbb{R}$ is a function and $\vec{v} : \mathbb{R}^3 \to \mathbb{R}^3$ is a vector field, then the following operators from vector calculus illustrated in Figure 15.1 are worth remembering:

Name	Notation	Definition
Gradient	∇f	$\left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}\right)$
Divergence	$\nabla \cdot \vec{v}$	$\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3}$
Curl	$\nabla \times \vec{v}$	$\left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}\right)$
Laplacian	$\nabla^2 f$	$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \frac{\partial^2 f}{\partial x_3^2}$

For PDEs involving fluids, electrodynamics, and other physical quantities, by convention we think of the derivatives above as acting on the *spatial* variables (x, y, z) rather than the time variable t. For instance, the gradient of a function $f : (x, y, z; t) \to \mathbb{R}$ will be written $\nabla f \equiv (\partial f/\partial x, \partial f/\partial y, \partial f/\partial z)$; the partial derivative in time $\partial f/\partial t$ is treated separately.

Example 15.1 (Fluid simulation)**.** The flow of fluids and smoke is governed by the *Navier-Stokes equations*, a system of PDEs in many variables. Suppose a fluid is moving in a region $\Omega \subseteq \mathbb{R}^3$. We define the following quantities:

$t \in [0, \infty)$	Time
$\vec{v}(t) : \Omega \to \mathbb{R}^3$	Velocity
$p(t) : \Omega \to \mathbb{R}$	Pressure
$\vec{f}(t) : \Omega \to \mathbb{R}^3$	External forces (e.g. gravity)



Boundary conditions (on $\partial\Omega$) Laplace solution (on Ω)

FIGURE 15.2 Laplace's equation takes a function on the boundary $\partial\Omega$ of a domain $\Omega \subseteq \mathbb{R}^2$ (left) and interpolates it to the interior of Ω as smoothly as possible (right).

If the fluid has fixed viscosity μ and density ρ, then the (incompressible) Navier-Stokes equations state:

$$\rho \cdot \left(\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} \right) = -\nabla p + \mu \nabla^2 \vec{v} + \vec{f} \qquad \text{with} \qquad \nabla \cdot \vec{v} = 0.$$

This system of equations determines the time dynamics of fluid motion and can be constructed by applying Newton's second law to tracking "particles" of fluid. Its statement, however, involves not only derivatives in time $\partial/\partial t$ but also derivatives in space ∇, making it a PDE.

Example 15.2 (Maxwell's equations)**.** Maxwell's equations determine the interaction between electric fields $\vec{E}$ and magnetic fields $\vec{B}$ over time. As with the Navier-Stokes equations, we think of the gradient, divergence, and curl operators as taking partial derivatives in space (x, y, z) and not time t. Then, in a vacuum Maxwell's system (in "strong" form) can be written:

$$\text{Gauss's law for electric fields: } \nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

$$\text{Gauss's law for magnetism: } \nabla \cdot \vec{B} = 0$$

$$\text{Faraday's law: } \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\text{Ampère's law: } \nabla \times \vec{B} = \mu_0 \left(\vec{J} + \varepsilon_0 \frac{\partial \vec{E}}{\partial t} \right)$$

Here, ε_0 and μ_0 are physical constants and $\vec{J}$ encodes the density of electrical current. Just like the Navier-Stokes equations, Maxwell's equations relate derivatives of physical quantities in time t to their derivatives in space (given by curl and divergence terms).

Example 15.3 (Laplace's equation)**.** Suppose Ω is a domain in $\mathbb{R}^2$ with boundary $\partial\Omega$ and that we are given a function $g : \partial\Omega \to \mathbb{R}$, illustrated in Figure 15.2. We may wish to interpolate g to the interior of Ω as smoothly as possible. When Ω is an irregular shape, however, our strategies for interpolation from Chapter 12 can break down.

Take $f(\vec{x}) : \Omega \to \mathbb{R}$ to be an interpolating function satisfying $f(\vec{x}) = g(\vec{x})$ for all

$\vec{x} \in \partial\Omega$. Then, one metric for evaluating the quality of f as a smooth interpolant is to define an energy functional:

$$E[f] = \int_\Omega \|\nabla f(\vec{x})\|_2^2 \, d\vec{x}$$

$E[f]$ measures the "total derivative" of f measured by taking the norm of its gradient and integrating this quantity over all of Ω. Wildly fluctuating functions f will have high values of $E[f]$ since the slope ∇f will be large in many places; smooth and low-frequency functions f, on the other hand, will have small $E[f]$ since their slope will be small everywhere. Here, the notation $E[\cdot]$ used here does not stand for "expectation" as it might in probability theory, but rather is an "energy" functional; it is standard notation in variational analysis.

We could ask that f interpolates g while being as smooth as possible in the interior of Ω using the following optimization:

$$\begin{aligned} &\text{minimize}_f \; E[f] \\ &\text{such that } f(\vec{x}) = g(\vec{x}) \, \forall x \in \partial\Omega. \end{aligned}$$

This setup *looks* like optimizations we have solved elsewhere, but now our unknown is a function f rather than a point in $\mathbb{R}^n$.

If f minimizes E subject to the boundary conditions, then $E[f+h] \geq E[f]$ for all functions $h(\vec{x})$ with $h(\vec{x}) = 0$ for all $\vec{x} \in \partial\Omega$. This statement is true even for small perturbations $E[f + \varepsilon h]$ as $\varepsilon \to 0$. Dividing by ε and taking the limit as $\varepsilon \to 0$, we must have $\frac{d}{d\varepsilon} E[f + \varepsilon h]|_{\varepsilon=0} = 0$; this expression is akin to setting directional derivatives of a function equal to zero to find its minima. We can simplify:

$$\begin{aligned} E[f + \varepsilon h] &= \int_\Omega \|\nabla f(\vec{x}) + \varepsilon \nabla h(\vec{x})\|_2^2 \, d\vec{x} \\ &= \int_\Omega (\|\nabla f(\vec{x})\|_2^2 + 2\varepsilon \nabla f(\vec{x}) \cdot \nabla h(\vec{x}) + \varepsilon^2 \|\nabla h(\vec{x})\|_2^2) \, d\vec{x} \end{aligned}$$

Differentiating with respect to ε shows:

$$\begin{aligned} \frac{d}{d\varepsilon} E[f + \varepsilon h] &= \int_\Omega (2\nabla f(\vec{x}) \cdot \nabla h(\vec{x}) + 2\varepsilon \|\nabla h(\vec{x})\|_2^2) \, d\vec{x} \\ \implies \frac{d}{d\varepsilon} E[f + \varepsilon h]|_{\varepsilon=0} &= 2 \int_\Omega [\nabla f(\vec{x}) \cdot \nabla h(\vec{x})] \, d\vec{x} \end{aligned}$$

Then, applying integration by parts and recalling that h is zero on $\partial\Omega$, we have:

$$\frac{d}{d\varepsilon} E[f + \varepsilon h]|_{\varepsilon=0} = -2 \int_\Omega h(\vec{x}) \nabla^2 f(\vec{x}) \, d\vec{x}$$

This expression must equal zero for *all* perturbations h that are zero on $\partial\Omega$. Hence, we must have $\nabla^2 f(\vec{x}) = 0$ for all $\vec{x} \in \Omega \backslash \partial\Omega$ (a formal proof is outside of the scope of our discussion).

We have shown that the boundary interpolation problem above amounts to solving the following PDE:

$$\begin{aligned} \nabla^2 f(\vec{x}) &= 0 \; \forall \vec{x} \in \Omega \backslash \partial\Omega \\ f(\vec{x}) &= g(\vec{x}) \; \forall \vec{x} \in \partial\Omega \end{aligned}$$

This PDE is known as *Laplace's equation*.

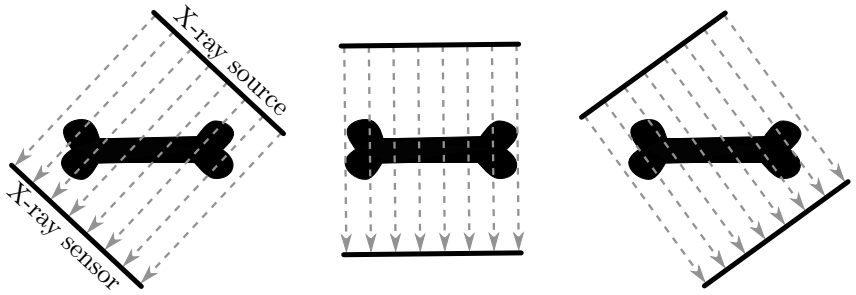


FIGURE 15.3 A CT scanner passes x-rays through an object; sensors on the other side collect the energy that made it through, giving the integrated density of the object along the x-ray path. Placing the source and sensor in different rotated poses allows for reconstruction of the pointwise density function.

Example 15.4 (X-ray computerized tomography)**.** Computerized tomography (CT) technology uses x-rays to see inside an object without cutting through it. The basic model is shown in Figure 15.3. Essentially, by passing x-rays through an object, the density of the object integrated along the x-ray path can be sensed by collecting the proportion that makes it through to the other side.

Suppose the density of an object is given by a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$. Then, for any two points $\vec{x}, \vec{y} \in \mathbb{R}^3$, we can think of a CT scanner abstractly as a device that can sense the integral u of ρ along the line connecting $\vec{x}$ and $\vec{y}$:

$$u(\vec{x}, \vec{y}) \equiv \int_{-\infty}^{\infty} \rho(t\vec{x} + (1-t)\vec{y})\, dt.$$

The function $u : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^+$ is known as the *Radon transform* of ρ.

Suppose we take a second derivative of u in an $\vec{x}$ and then a $\vec{y}$ coordinate:

$$\begin{aligned}
\frac{\partial}{\partial x_i} u(\vec{x}, \vec{y}) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial x_i} \rho(t\vec{x} + (1-t)\vec{y})\, dt \text{ by definition of } u \\
&= \int_{-\infty}^{\infty} t\vec{e}_i \cdot \nabla\rho(t\vec{x} + (1-t)\vec{y})\, dt \\
\implies \frac{\partial^2}{\partial y_j \partial x_i} u(\vec{x}, \vec{y}) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial y_j} t\vec{e}_i \cdot \nabla\rho(t\vec{x} + (1-t)\vec{y})\, dt \\
&= \int_{-\infty}^{\infty} t(1-t)\vec{e}_i^{\top} H_\rho(t\vec{x} + (1-t)\vec{y})\vec{e}_j\, dt \text{ for Hessian } H_\rho \text{ of } \rho.
\end{aligned}$$

An identical set of steps shows that the derivative $\frac{\partial^2}{\partial x_j \partial y_i} u$ equals the same expression after applying symmetry of H_ρ. That is, u satisfies the following relationship:

$$\frac{\partial^2 u}{\partial y_j \partial x_i} = \frac{\partial^2 u}{\partial x_j \partial y_i}$$

This equality, known as the Fritz John equation [63], gives information about u without involving the unknown density function ρ. In a computational context, it can be used to fill in data missing from incomplete x-ray scans or to smooth data from a potentially noisy x-ray sensor before reconstructing ρ.

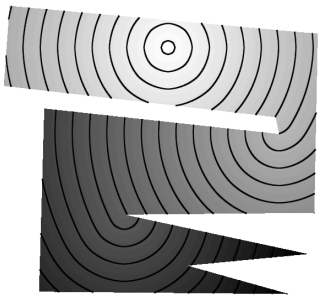


FIGURE 15.4 Shortest-path distances constrained to move within the interior of a non-convex shape have to wrap around corners; level sets of the distance function (shown as black lines) are no longer circles beyond these corner points.

Example 15.5 (Eikonal equation)**.** Suppose Ω is a closed region in $\mathbb{R}^n$. For a fixed point $\vec{x}_0 \in \Omega$, we might wish to find a function $d(\vec{x}) : \Omega \to \mathbb{R}^+$ measuring the length of the shortest path from $\vec{x}_0$ to $\vec{x}$ restricted to move only within Ω. When Ω is convex, we can write d in closed form:

$$d(\vec{x}) = \|\vec{x} - \vec{x}_0\|_2.$$

As illustrated in Figure 15.4, however, if Ω is non-convex or is a complicated domain like a surface, these distance functions become more challenging to compute. Solving for d, however, is a critical step for tasks like planning paths of robots by minimizing the distance they travel while avoiding obstacles marked on a map.

If Ω is non-convex, away from singularities the function $d(\vec{x})$ still satisfies a derivative condition known as the *eikonal equation*:

$$\|\nabla d\|_2 = 1.$$

Intuitively, this PDE states that a distance function should have unit rate of change everywhere; as a sanity check, this is certainly true for the absolute value function $|x - x_0|$ in one dimension, which measures the distance along the real line between x_0 and x. This equation is nonlinear in the derivative ∇d, making it a particularly challenging problem.

Specialized algorithms known as *fast marching methods* and *fast sweeping methods* estimate $d(\vec{x})$ over all of Ω by integrating the eikonal equation. Interestingly but unsurprisingly, many algorithms for approximating solutions to the eikonal equation have structure similar to Dijkstra's algorithm for computing shortest paths along graphs. See problem 15.8 for one example.

Example 15.6 (Harmonic analysis)**.** Different objects respond differently to vibrations, and in large part these responses are functions of the *geometry* of the objects. For example, cellos and pianos can play the same note, but even an inexperienced listener can distinguish between the sounds they make.

From a mathematical standpoint, we can take $\Omega \subseteq \mathbb{R}^3$ to be a shape represented either as a surface or a volume. If we clamp the edges of the shape, then its *frequency spectrum* is given by eigenvalues coming from the following problem:

$$\begin{aligned} \nabla^2 \phi &= \lambda \phi \\ \phi(\vec{x}) &= 0 \ \forall \vec{x} \in \partial\Omega, \end{aligned}$$

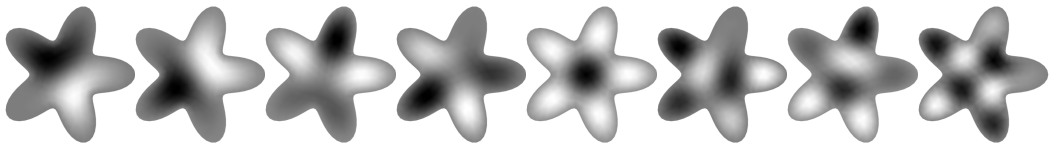


FIGURE 15.5 The first eight eigenfunctions ϕ_i of the Laplacian operator of the domain Ω from Figure 15.2, which satisfy $\nabla^2\phi_i = \lambda_i\phi_i$ in order of increasing frequency; we omit ϕ_1, which is the constant function with $\lambda = 0$.

where ∇^2 is the Laplacian of Ω and $\partial\Omega$ is the boundary of Ω. Figure 15.5 shows examples of these functions on a two-dimensional domain Ω.

Relating to the one-dimensional theory of waves, $\sin kx$ solves this problem when Ω is the interval $[0, 2\pi]$ and $k \in \mathbb{Z}$. To check, the Laplacian in one dimension is $\partial^2/\partial x^2$, and thus

$$\begin{aligned}\frac{\partial^2}{\partial x^2}\sin kx &= \frac{\partial}{\partial x}k\cos kx \\ &= -k^2\sin kx \\ \sin(k\cdot 0) &= 0 \\ \sin(k\cdot 2\pi) &= 0.\end{aligned}$$

That is, the eigenfunctions are $\sin kx$ with eigenvalues $-k^2$.

15.2 STATEMENT AND STRUCTURE OF PDES

Vocabulary used to describe PDEs is extensive, and each class of PDEs has substantially different properties from the others in terms of solvability, theoretical understanding of solutions, and discretization challenges. Our main focus eventually will be on developing algorithms to a few common tasks rather than introducing the general theory of continuous or discretized PDE, but it is worth acknowledging the rich expressive possibilities—and accompanying theoretical challenges—that come with using PDE language to describe numerical problems.

Following standard notation, for our subsequent development, we will assume that our unknown is some function $u(\vec{x})$. For ease of notation, we will use subscript notation to denote partial derivatives:

$$u_x \equiv \frac{\partial u}{\partial x}, \qquad u_y \equiv \frac{\partial u}{\partial y}, \qquad u_{xy} \equiv \frac{\partial^2 u}{\partial x \partial y},$$

and so on.

15.2.1 Properties of PDEs

Just as ODEs couple the time derivatives of a function, PDEs typically are stated as relationships between two or more partial derivatives of u. By examining the algebraic form of a PDE, we can check if it has any of a number of properties, including the following:

- Homogeneous (e.g. $x^2u_{xx} + u_{xy} - u_y + u = 0$): The PDE can be written using linear

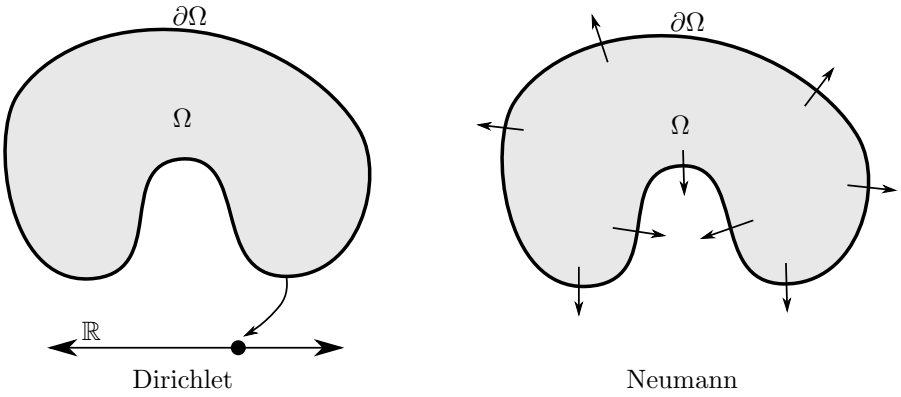


FIGURE 15.6 Dirichlet boundary conditions prescribe the values of the unknown function u on the boundary $\partial\Omega$ of the domain Ω, while Neumann conditions prescribe the derivative of u orthogonal to $\partial\Omega$.

combinations of u and its derivatives; the coefficients can be scalar values or functions of the independent variables. The equation can be nonlinear in the independent variables (x and y in our example).

- Linear (e.g. $u_{xx} - yu_{yy} + u = xy^2$): Similar to homogeneous PDE, but potentially with a nonzero (inhomogeneous) right-hand side built from scalars or the dependent variables. Contrastingly, PDEs like the eikonal equation (or $u_{xx}^2 = u_{xy}$) are considered nonlinear because they are nonlinear in u.
- Quasi-linear (e.g. $u_{xy}+2u_{xx}+u_y^2+u_x^2 = y$): The statement is linear in the highest-order derivatives of u.
- Constant-coefficient (e.g. $u_{xx}+3u_y = u_z$): The coefficients of u and its derivatives are not functions of the independent variables.

One potentially surprising observation about the properties above is that they are more concerned with the role of u than that of the independent variables like x, y, and z. For instance, the definition of a "linear" PDE allows u to have coefficients that are nonlinear functions of these variables. While this may make the PDE appear nonlinear, it is still linear in the unknowns, which is the distinguishing factor.

The *order* of a PDE is the order of its highest derivative. Most of the PDEs we consider in this chapter are second-order and already present considerable numerical challenges. Methods analogous to reduction of ODEs to first order (§14.2.1) can be carried out but do not provide as much benefit for solving PDE.

15.2.2 Boundary Conditions

ODEs typically are considered *initial-value problems*, because given a configuration that is known at the initial time $t = 0$, they evolve the state forward indefinitely. With few exceptions, the user does not have to provide information about the state for $t > 0$.

PDE problems also can be *boundary-value problems* rather than or in addition to being initial value problems. Most PDEs require information about behavior at the boundary of the domain of all the variables of the domain. For instance, Laplace's equation as introduced

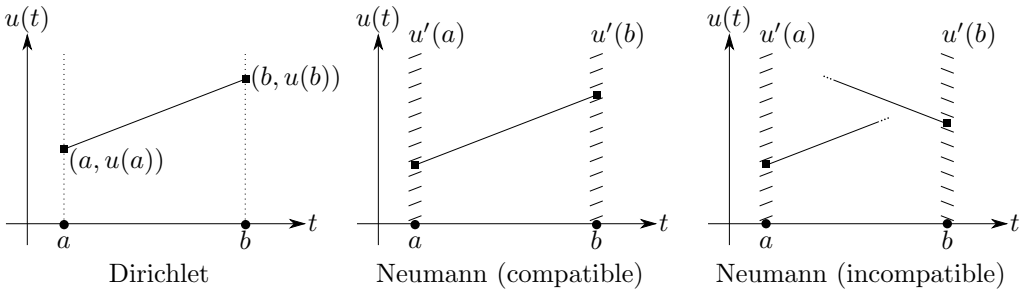


FIGURE 15.7 Boundary conditions for the PDE $u_{tt} = 0$ from Example 15.7.

in Example 15.3 requires fixed values on the boundary $\partial\Omega$ of the domain Ω. Similarly, the heat equation used to simulate conductive material like metals admits a number of possible boundary conditions, corresponding to whether the material is attached to a heat source or dispersing heat energy into the surrounding space.

If the unknown of a PDE is a function $u : \Omega \to \mathbb{R}$ for some domain $\Omega \subseteq \mathbb{R}^n$, typical boundary conditions include the following:

- *Dirichlet conditions* directly specify the values of $u(\vec{x})$ for all $\vec{x} \in \partial\Omega$.
- *Neumann conditions* specify the derivative of $u(\vec{x})$ in the direction orthogonal to $\partial\Omega$.
- *Mixed* or *Robin conditions* specify a relationship between the value and normal derivatives of $u(\vec{x})$ on $\partial\Omega$.

The first two choices are illustrated in Figure 15.6.

Improperly encoding boundary conditions is a subtle oversight that creeps into countless discretizations of PDEs. There are many sources of confusion that explain this common issue. Different discretizations of the same boundary conditions can yield qualitatively different outputs from a PDE solver if they are expressed improperly. Indeed, some boundary conditions are not realizable even in theory, as illustrated in the example below.

Example 15.7 (Boundary conditions in one dimension)**.** Suppose we are solving the following PDE (more precisely an ODE, although the distinction here is not relevant) in one variable t over the interval $\Omega = [a, b]$:

$$u_{tt} = 0.$$

From one-variable calculus, we know that solutions must take the form $u(t) = \alpha t + \beta$, a line.

Consider the effects of assorted choices of boundary conditions on $\partial\Omega = \{a, b\}$, illustrated in Figure 15.7:

- Dirichlet conditions specify the values $u(a)$ and $u(b)$ directly. There is a unique line that goes through any pair of points $(a, u(a))$ and $(b, u(b))$, so a solution to the PDE always exists in this case.
- Neumann conditions specify $u'(a)$ and $u'(b)$. From our general form of $u(t)$, we know $u'(t) = \alpha$, reflecting the fact that lines have constant slope. In this case, Neumann conditions requiring that $u'(a) \neq u'(b)$ are *incompatible* with the PDE itself.

15.3 MODEL EQUATIONS

In §14.2.3, we demonstrated properties of ODEs and their integrators by examining the model equation $y' = ay$. We can pursue a similar analytical technique for PDEs, although we will have to separate into multiple special cases to cover the qualitative phenomena of interest.

We will focus on the linear, constant-coefficient, homogeneous case. As mentioned in §15.2.1, the non-constant coefficient and inhomogeneous cases often have similar qualitative behavior, and nonlinear PDEs require special consideration beyond the scope of our discussion. We furthermore will study *second-order* systems, that is, systems containing at most the second derivative of u. While the model ODE $y' = ay$ is first-order, a reasonable model PDE needs at least two derivatives to show how derivatives in different directions interact.

Linear, constant-coefficient, homogeneous second-order PDEs have the following general form, for unknown function $u : \mathbb{R}^n \to \mathbb{R}$:

$$\sum_{ij} a_{ij} \frac{\partial u}{\partial x_i \partial x_j} + \sum_i b_i \frac{\partial u}{\partial x_i} + cu = 0.$$

To simplify notation, we can define a formal "gradient operator" as the vector of derivatives:

$$\nabla \equiv \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n} \right).$$

Expressions like ∇f, $\nabla \cdot \vec{v}$, and $\nabla \times \vec{v}$ agree with the definitions of gradients, divergence, and curl on $\mathbb{R}^3$ using this formal definition of ∇. In this notation, the model PDE takes a matrix form:

$$(\nabla^\top A \nabla + \nabla \cdot \vec{b} + c)u = 0.$$

The operator $\nabla^\top A \nabla + \nabla \cdot \vec{b} + c$ acting on u abstractly looks like a quadratic form in ∇ as a vector; since partial derivatives commute, we can assume A is symmetric.

The definiteness of A determines the *class* of the model PDE, just as the definiteness of a matrix determines the convexity of its associated quadratic form. Four cases bring about qualitatively different behavior for u:

- If A is *positive or negative definite*, system is *elliptic*.
- If A is *positive or negative semidefinite*, the system is *parabolic*.
- If A has only one eigenvalue of different sign from the rest, the system is *hyperbolic*.
- If A satisfies none of the criteria, the system is *ultrahyperbolic*.

These criteria are listed approximately in order of the difficulty level of solving each type of equation. We consider the first three cases below and provide examples of corresponding behavior by specifying different matrices A; ultrahyperbolic equations do not appear as often in practice and require highly specialized solution techniques.

15.3.1 Elliptic PDEs

Positive definite matrices admitted efficient linear solvers like Cholesky decomposition and conjugate gradients that did not hold in the general case. Similarly, elliptic PDEs, for which A is positive definite, have strong structure that makes them the most straightforward equations to characterize and solve, both theoretically and computationally.

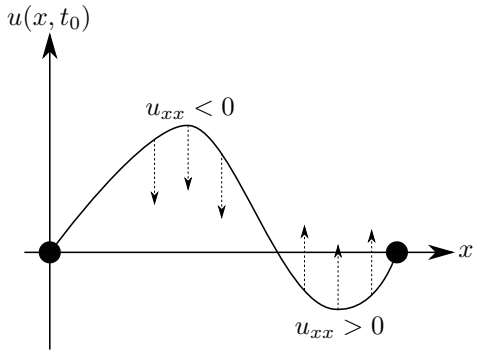


FIGURE 15.8 The heat equation in one variable $u_t = \alpha u_{xx}$ decreases u over time where it is curved down and increase u over time where u is curved up, as measured using the second derivative in space u_{xx}. Here, we show a solution of the heat equation $u(x, t)$ at a fixed time t_0; the arrows indicate how values of u will change as t advances.

The model elliptic PDE is the *Laplace equation*, given by $\nabla^2 u = 0$ as in Example 15.3. For instance, in two variables the Laplace equation becomes

$$u_{xx} + u_{yy} = 0.$$

Figure 15.2 illustrated a solution of the Laplace equation, which essentially interpolates information from the boundary of the domain of u to its interior.

Elliptic equations are well-understood theoretically and come with strong properties characterizing their behavior. Of particular importance is *elliptic regularity*, which states that solutions of elliptic PDEs automatically are differentiable to higher order than their building blocks. Physically, elliptic equations compute stable equilbria like the rest pose of a stretched rubber sheet, which naturally resists kinks and other irregularities.

15.3.2 Parabolic PDEs

Positive *semi*definite linear systems are only marginally more difficult to deal with than positive definite ones, at least if their null spaces are known and relatively small. In particular, positive semidefinite matrices have null spaces that prevent them from being invertible, but orthogonally to the null space they behave identically to definite matrices. In PDE, these systems correspond to *parabolic* equations, for which A is positive semidefinite.

The heat equation is the model parabolic PDE. Suppose $u_0(x, y)$ is a fixed distribution of temperature in some region $\Omega \subseteq \mathbb{R}^2$ at time $t = 0$. Then, the heat equation determines how heat diffuses over time $t > 0$ as a function $u(t; x, y)$:

$$u_t = \alpha(u_{xx} + u_{yy}),$$

where $\alpha > 0$. If $\nabla = (\partial/\partial x, \partial/\partial y)$, the heat equation can be written $u_t = \alpha \nabla^2 u$. The indefinite derivative is in time t, making the equation parabolic rather than elliptic.

Figure 15.8 provides a phenomenological interpretation of the heat equation in one variable. We can think of the second derivative $\nabla^2 u$ as measuring the convexity of u. The heat equation increases u with time when its value is "cupped" upward, and decreases u

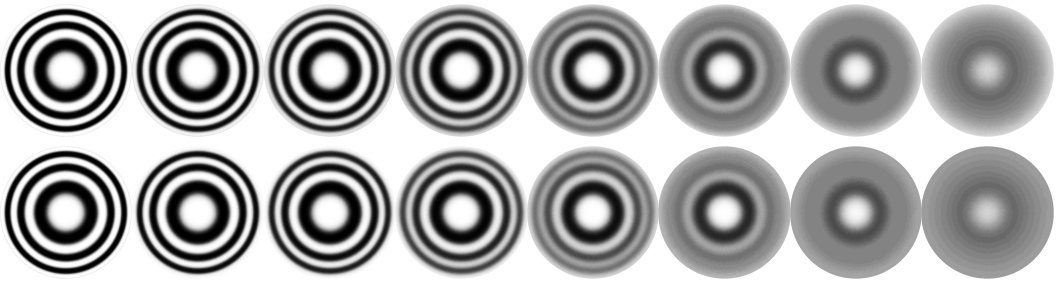


FIGURE 15.9 Solution to the heat equation $u_t = u_{xx} + u_{yy}$ on the unit circle with Dirichlet (top) and Neumann (bottom) boundary conditions. Solutions are colored from -1 (black) to 1 (white).

otherwise. This negative feedback is stable and leads to equilibrium as $t \to \infty$. An example solution to the heat equation is shown in Figure 15.9.

The corresponding second-order term matrix A for the heat equation is:

$$A = \begin{array}{c} \\ t \\ x \\ y \end{array} \begin{array}{c} \begin{array}{ccc} t & x & y \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

The heat equation is parabolic since this matrix has eigenvalues 0, 1, and 1, since the second derivative with respect to t does not appear.

There are two boundary conditions needed for the heat equation, both of which have physical interpretations:

- The distribution of heat $u(0; x, y) \equiv u_0(x, y)$ at time $t = 0$ at all points $(x, y) \in \Omega$
- Behavior of u when $t > 0$ at points $(x, y) \in \partial\Omega$. Dirichlet conditions provide $u(t; x, y)$ for all $t \geq 0$ and $(x, y) \in \partial\Omega$, e.g. if Ω is a piece of foil sitting next to a heat source like an oven whose temperature is controlled externally. Contrastingly, Neumann conditions specify the derivative of f in the direction normal to the boundary $\partial\Omega$; they correspond to fixing the *flux* of heat out of Ω caused by different types of insulation.

15.3.3 Hyperbolic PDEs

The final model equation is the wave equation, corresponding to the indefinite matrix case:

$$u_{tt} = c^2(u_{xx} + u_{yy})$$

The wave equation is hyperbolic because the second derivative in time t has opposite sign from the two spatial derivatives. This equation determines the motion of waves across an elastic medium like a rubber sheet. It can be derived by applying Newton's second law to points on a piece of elastic, where x and y are positions on the sheet and $u(t; x, y)$ is the height of the piece of elastic at time t.

Figure 15.10 illustrates a solution of the wave equation with Dirichlet boundary conditions; these boundary conditions correspond to the vibrations of a drum whose outer

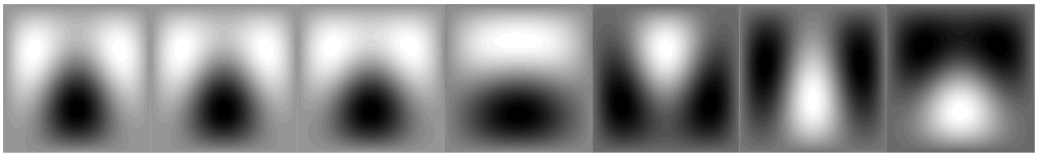


FIGURE 15.10 The wave equation on a square with Dirichlet boundary conditions; time is sampled evenly and progresses left-to-right. Color is proportional to the height of the wave, from -1 (black) to 1 (white).

boundary is fixed. As illustrated in the example, wave behavior contrasts considerably with heat diffusion in that as $t \to \infty$ the energy of the system does not disperse; waves can bounce back and forth across a domain indefinitely. For this reason, implicit integration strategies may not be appropriate for integrating hyperbolic PDEs because they tend to damp out motion.

Boundary conditions for the wave equation are similar to those of the heat equation, but now we must specify both $u(0; x, y)$ and $u_t(0; x, y)$ at time zero:

- The conditions at $t = 0$ specify the position and velocity of the wave at the start time.
- Boundary conditions on $\partial\Omega$ determine what happens at the ends of the material. Dirichlet conditions correspond to fixing the sides of the wave, e.g. plucking a cello string that is held flat at its two ends on the instrument. Neumann conditions correspond to leaving the ends of the wave untouched, like the end of a whip.

15.4 REPRESENTING DERIVATIVE OPERATORS

A key intuition that underlies many numerical techniques for PDEs is the following:

Derivatives act on functions in the same way that sparse matrices act on vectors.

Our choice of notation reflects this parallel: The derivative $d/dx[f(x)]$ looks like the product of an operator d/dx and a function f; defining $\nabla \equiv (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ even stacked derivative operators into a vector.

Formally, differentiation is a *linear operator* like matrix multiplication, since for all smooth functions $f, g : \mathbb{R} \to \mathbb{R}$ and scalars $a, b \in \mathbb{R}$,

$$\frac{d}{dx}(af(x) + bg(x)) = a\frac{d}{dx}f(x) + b\frac{d}{dx}g(x).$$

The derivatives act on functions, which can be thought of as points in an infinite-dimensional vector space. Many arguments from Chapter 0 and elsewhere regarding the linear algebra of matrices extend to this case, providing conditions for invertibility, symmetry, and so on of these abstract operators.

Nearly all techniques for solving linear PDEs make this analogy concrete. For example, recall the model equation $(\nabla^\top A\nabla + \nabla \cdot \vec{b} + c)u = 0$ subject to Dirichlet boundary conditions $u|_{\partial\Omega} = u_0$ for some fixed function u_0. We can define an operator $R_{\partial\Omega} : C^\infty(\Omega) \to C^\infty(\partial\Omega)$, that is, an operator taking functions on Ω and returning functions on its boundary $\partial\Omega$,

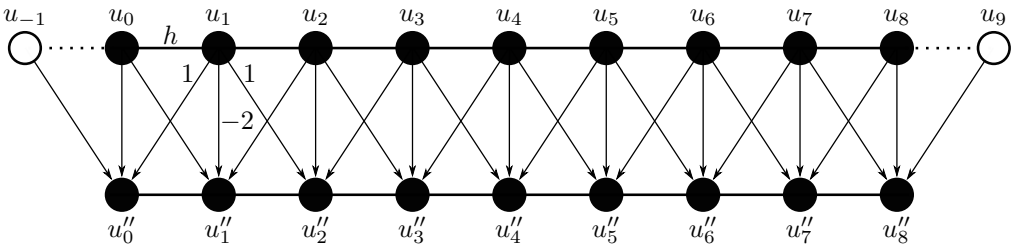


FIGURE 15.11 The one-dimensional finite difference Laplacian operator L takes samples u_i of a function $u(x)$ and returns an approximation of u'' at the same grid points by combining neighboring values using weights (1)—(−2)—(1); here $u(x)$ is approximated using nine samples $u_0, \ldots, u_8$. Boundary conditions are needed to deal with the unrepresented quantities at the white endpoints.

by restriction: $[R_{\partial\Omega}u](\vec{x}) \equiv u(\vec{x})$ for all $\vec{x} \in \partial\Omega$. Then, the model PDE and its boundary conditions can be combined in matrix-like notation:

$$\begin{pmatrix} (\nabla^\top A \nabla + \nabla \cdot \vec{b} + c) \\ R_{\partial\Omega} \end{pmatrix} u = \begin{pmatrix} 0 \\ u_0 \end{pmatrix}.$$

More compactly, we wish to solve $Mu = w$ where M is linear. If we discretize M as a matrix, then recovering the solution u of the original equation is as easy as writing

$$\text{“}u = M^{-1}w.\text{”}$$

Many discretizations exist M and u, often derived from the discretizations of derivatives introduced in §13.3. While each has subtle advantages, disadvantages, and conditions for effectiveness or convergence, in this section we provide constructions and high-level themes from a few popular techniques. Realistically speaking, a legitimate and often-applied technique for finding the best discretization for a given application is to try a few and check empirically which is the most effective.

15.4.1 Finite Differences

Consider a function $u(x)$ on $[0, 1]$. Using the methods from Chapter 13, we can approximate the second derivative $u''(x)$ as

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^2).$$

In the course of solving a PDE in u, assume $u(x)$ is discretized using $n+1$ evenly-spaced samples $u_0, u_1, \ldots, u_n$, as in Figure 15.11, and take h to be the spacing between samples, satisfying $h = 1/n$. Applying our formula above provides an approximation of u'' at each grid point:

$$u_k'' \approx \frac{u_{k+1} - 2u_k + u_{k-1}}{h^2}$$

That is, the second derivative of a function on a grid of points can be estimated using the (1)—(−2)—(1) stencil illustrated in Figure 15.12.

Boundary conditions are needed to compute u_0'' and u_n'' since we have not included u_{-1} or u_{n+1} in our discretization. Keeping in mind that $u_0 = u(0)$ and $u_n = u(1)$, we can incorporate them as follows:

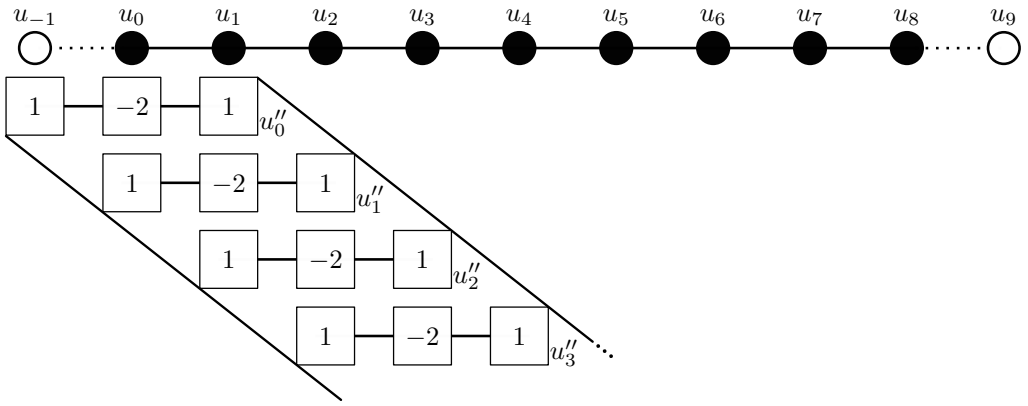


FIGURE 15.12 The one-dimensional finite difference Laplacian can be thought of as dragging a (1)—(−2)—(1) stencil across the domain.

- Dirichlet: $u_{-1} \equiv u_{n+1} = 0$, that is, fix the value of u beyond the endpoints to be zero
- Neumann: $u_{-1} = u_0$ and $u_{n+1} = u_n$, encoding the condition $u'(0) = u'(1) = 0$
- Periodic: $u_{-1} = u_n$ and $u_{n+1} = u_0$, making the identification $u(0) = u(1)$

Suppose we stack the samples u_k into a vector $\vec{u} \in \mathbb{R}^{n+1}$ and the samples u_k'' into a second vector $\vec{w} \in \mathbb{R}^{n+1}$. Then, our construction above shows that $h^2\vec{w} = L\vec{u}$, where L is one of the choices below:

Dirichlet

$$\begin{pmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2 & 1 & \\ & & & 1 & -2 \end{pmatrix}$$

Neumann

$$\begin{pmatrix} -1 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2 & 1 & \\ & & & 1 & -1 \end{pmatrix}$$

Periodic

$$\begin{pmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2 & 1 & \\ 1 & & & 1 & -2 \end{pmatrix}$$

The matrix L can be thought of as a discretized version of the operator $\frac{d^2}{dx^2}$ acting on $\vec{u} \in \mathbb{R}^{n+1}$ rather than functions $u : [0,1] \to \mathbb{R}$.

In two dimensions, we can write a similar approximation for the Laplacian $\nabla^2 u$ for $u : [0,1] \times [0,1] \to \mathbb{R}$. Now, we sample using a grid of values shown in Figure 15.13. In this case, $\nabla^2 u = u_{xx} + u_{yy}$, so we sum up x and y second derivatives constructed in the one-dimensional example above. If we number our samples as $u_{k,\ell} \equiv u(kh, \ell h)$, then our formula for the Laplacian of u is:

$$(\nabla^2 u)_{k,\ell} \approx \frac{u_{(k-1),\ell} + u_{k,(\ell-1)} + u_{(k+1),\ell} + u_{k,(\ell+1)} - 4u_{k,\ell}}{h^2}.$$

This approximation implies a (1)—(−4)—(1) stencil over a 3×3 box. If we once again combine our samples of u and ∇u into $\vec{u}$ and $\vec{w}$, resp., then $h^2\vec{w} = L_2\vec{y}$ where L_2 comes from the stencil we derived. This two-dimensional grid Laplacian L_2 appears in many image processing applications, where (k, ℓ) is used to index pixels on an image.

Regardless of dimension, given a discretization of the domain and a Laplacian matrix L, we can approximate solutions of elliptic PDEs using linear systems of equations. Consider the Poisson equation $\nabla^2 u = w$. After applying our discretization, given a sampling $\vec{w}$ of

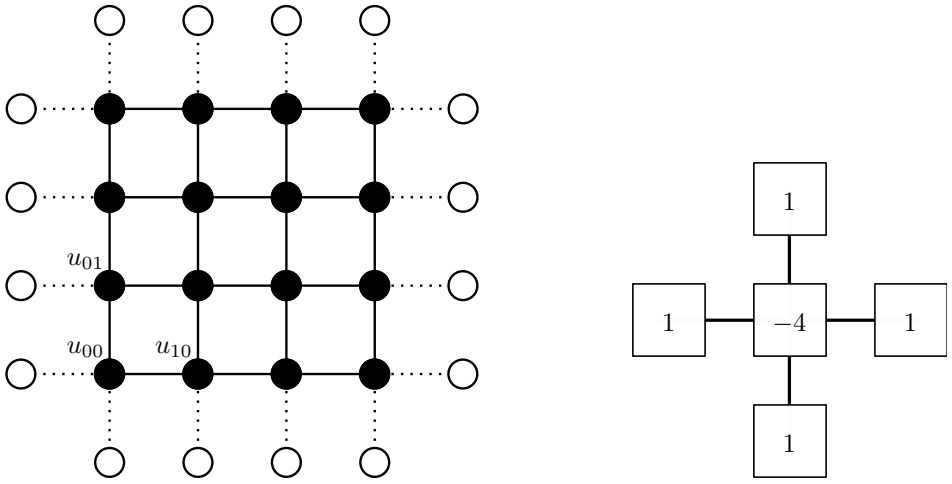


FIGURE 15.13 For functions $u(x, y)$ discretized on a two-dimensional grid (left), the Laplacian L_2 has a (1)—(−4)—(1) stencil.

$w(\vec{x})$, we can obtain an approximation $\vec{u}$ of the solution by solving the system $L\vec{u} = h^2\vec{w}$ for $\vec{u}$.

This approach can be extended to inhomogeneous boundary conditions. For example, if we wish to solve $\nabla^2 u = w$ on a two-dimensional grid subject to Dirichlet conditions prescribed by a function u_0, we could do so by solving the following linear system of equations for $\vec{u}$:

$$\begin{aligned} u_{k,\ell} &= u_0(kh, lh) && \text{when } k \in \{0, n\} \text{ or } \ell \in \{0, n\} \\ u_{(k-1),\ell} + u_{k,(\ell-1)} + u_{(k+1),\ell} + u_{k,(\ell+1)} - 4u_{k,\ell} &= 0 && \text{otherwise} \end{aligned}$$

This system of equations uses the 3×3 Laplacian stencil for vertices in the interior of $[0, 1]^2$ while explicitly fixing the values of u on the boundary.

These discretizations exemplify the *finite differences* method of discretizing PDEs, usually when the domain can be approximated using a grid. Finite differences essentially treats the divided difference approximations from Chapter 13 as linear operators on grids of function values and then solves the resulting discrete system of equations.

Quoting results from Chapter 13 directly, however, comprises a serious breach of notation. When we write that an approximation of $u'(x)$ or $u''(x)$ holds to $O(h^k)$, we implicitly assume that $u(x)$ is sufficiently differentiable. Hence, what we need to show is that the result of solving systems like $L\vec{u} = h^2\vec{w}$ produces a $\vec{u}$ that actually approximates samples from a smooth function $u(x)$ rather than oscillating crazily. The following example shows that this issue is practical rather than theoretical, and that reasonable but non-convergent discretizations can fail catastrophically.

Example 15.8 (Lack of convergence)**.** Suppose we again sample a function $u(x)$ of one variable and wish to solve an equation that involves a first-order u' term. Interestingly, this task can be more challenging than solving second-order equations.

First, if we define u'_k as the forward difference $\frac{1}{h}(u_{k+1} - u_k)$, then we will be in the unnaturally asymmetric position of needing a boundary condition at u_n but not at u_0 as shown in Figure 15.14. Backward differences suffer from the reverse problem.

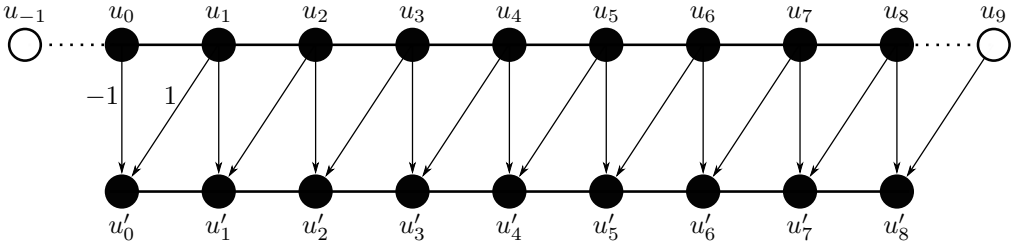


FIGURE 15.14 Forward differencing to approximate $u'(x)$ asymmetrically requires boundary conditions on the right but not the left.

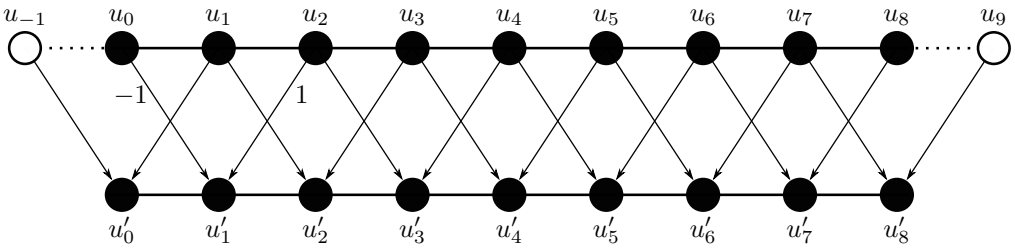


FIGURE 15.15 Centered differencing yields a symmetric approximation of $u'(x)$, but u'_k is not affected by the value of u_k using this formula.

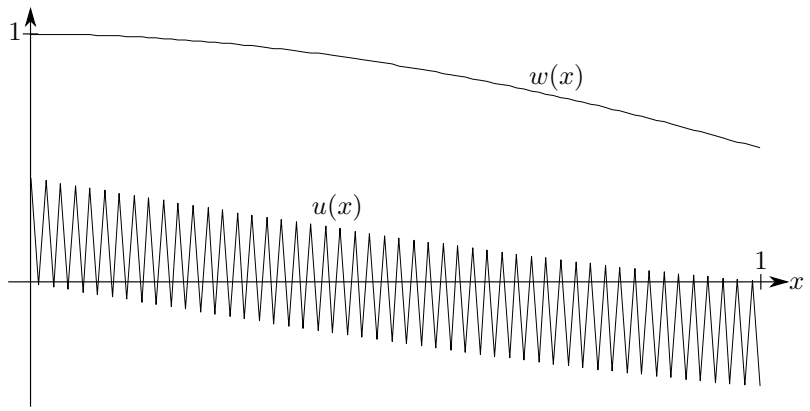


FIGURE 15.16 Solving $u'(x) = w(x)$ for $u(x)$ using a centered difference discretization suffers from the fencepost problem; odd- and even-indexed values of u have completely separate behavior. As more gridpoints are added in x, the resulting $u(x)$ does not converge to a smooth function, so $O(h^k)$ estimates of derivative quality do *not* apply.

We might attempt to solve this problem and simultaneously gain an order of accuracy by using the symmetric difference $u'_k \approx \frac{1}{2h}(u_{k+1} - u_{k-1})$, but this discretization suffers from a more subtle *fencepost problem* illustrated in Figure 15.15. In particular, this version of u'_k ignores the value of u_k itself and only looks at its neighbors u_{k-1} and u_{k+1}. This oversight means that u_k and u_ℓ are treated differently depending on whether k and ℓ are even or odd. Figure 15.16 shows the result of attempting to solve a numerical problem with this discretization; the result is non-differentiable.

As with the leapfrog integration algorithm in §14.4.2, one way to avoid these issues is to think of the derivatives as living on *half* gridpoints. In the one-dimensional case, this change corresponds to labeling the difference $\frac{1}{h}(y_{k+1} - y_k)$ as $y_{k+1/2}$. This technique of placing different derivatives on vertices, edges, and centers of grid cells is particularly common in fluid simulation, which maintains pressures, fluid velocities, and other physical quantities at locations suggested by the discretization.

15.4.2 Collocation

A challenge when working with finite differences is that we must justify that the end result "looks like" the theoretical solution we are seeking to approximate. That is, we have replaced a continuous unknown $u(\vec{x})$ with a sampled proxy on a grid but may inadvertently lose the connection to the continuous case in the process; Example 15.8 shows one example where a discretization is not convergent and hence yields unusable output. For this reason, many numerical PDE methods attempt to make the connection between continuous and discrete less subtle.

One way to link continuous and discrete models of PDE is to write $u(\vec{x})$ in a basis $\phi_1, \ldots, \phi_k$ as

$$u(\vec{x}) \approx \sum_{i=1}^{k} a_i \phi_i(\vec{x}).$$

This strategy should be familiar, as it underlies formulations we have considered in previous chapters for interpolation, quadrature, and differentiation. The philosophy here is to find coefficients $a_1, \ldots, a_k$ providing the *best possible approximation* of the solution to the continuous problem in the ϕ_i basis; as we add more functions ϕ_i to the basis, in many cases the approximation will converge to the theoretical solution, so long as the ϕ_i's eventually cover the relevant part of function space.

Perhaps the simplest method making use of this new construction is the *collocation* method. In the presence of k basis functions, this method samples k points $\vec{x}_1, \ldots, \vec{x}_k \in \Omega$ and requires that the PDE holds exactly at these locations. For example, if we wish to solve the Poisson equation $\nabla^2 u = w$, then for each $i \in \{1, \ldots, k\}$ we write the relationship:

$$w(\vec{x}_i) = \nabla^2 u(\vec{x}_i) = \sum_{j=1}^{k} a_j \nabla^2 \phi_j(\vec{x}_i)$$

This square linear system can be solved for the vector $\vec{a} \in \mathbb{R}^k$ of coefficients. It can be replaced with a least-squares problem if more than k points are sampled in Ω.

Collocation requires the choice of basis functions $\phi_1, \ldots, \phi_k$ and *collocation points* $\vec{x}_1, \ldots, \vec{x}_k$. Typical basis functions include full or piecewise polynomial functions and trigonometric functions. When the ϕ_i's are compactly-supported, that is, when $\phi_i(\vec{x}) = 0$ for most $\vec{x} \in \Omega$, the resulting matrix to be solved becomes sparse. The output of collocation is a set of coefficients rather than a set of function values as in finite differences. Since the basis

functions do not have to have any sort of grid structure, it is well-suited to non-rectangular domains, which can provide some challenge for finite differencing.

A drawback of collocation is that it does not deal with the behavior of the approximation $u(\vec{x})$ between the collocation points. Just as interpolating a polynomial through a set of sample points can lead to degenerate and in some cases highly-oscillatory behavior, the collocation method must be used with caution to avoid degenerate output, for instance by optimizing the choice of basis functions and collocation points. Another option is to use a method like finite elements, considered below, which integrates behavior of an approximation over more than one sample point at a time.

15.4.3 Finite Elements

Finite elements discretizations also makes use of basis functions but does so by examining integrated quantities rather than pointwise values of the unknown function $u(\vec{x})$. This type of discretization is relevant to simulating a wide variety of phenomena and remains a popular choice in a diverse set of fields including mechanical engineering, digital geometry processing, and cloth simulation.

As an example, suppose that $\Omega \subseteq \mathbb{R}^2$ is some region on the plane and that we wish to solve the Dirichlet equation $\nabla^2 u = 0$ in its interior. Take any other function $v(\vec{x})$, satisfying $v(\vec{y}) = 0$ for all $\vec{y} \in \partial\Omega$. If we solve the PDE successfully, the function $u(\vec{x})$ will satisfy the relationship

$$\int_\Omega v(\vec{x})\nabla^2 u(\vec{x})\,d\vec{x} = \int_\Omega v(\vec{x}) \cdot 0\,d\vec{x} = 0,$$

regardless of the choice of v.

We can define a bilinear operator $\langle u, v\rangle_{\nabla^2}$ as the integral

$$\langle u, v\rangle_{\nabla^2} \equiv \int_\Omega v(\vec{x})\nabla^2 u(\vec{x})\,d\vec{x}.$$

Any function $u(\vec{x})$ for which $\langle u, v\rangle_{\nabla^2} = 0$ for all reasonable $v : \Omega \to \mathbb{R}$ defined above is called a *weak solution* to the Dirichlet equation. The functions v are known as *test functions*.

A remarkable observation suggests that weak solutions to PDEs may exist even when a strong solution does not. In particular, when $v(\vec{x})$ is zero on $\partial\Omega$, the *divergence theorem* from multivariable calculus implies the following alternative form for $\langle u, v\rangle_{\nabla^2}$:

$$\langle u, v\rangle_{\nabla^2} = -\int_\Omega \nabla u(\vec{x}) \cdot \nabla v(\vec{x})\,d\vec{x}.$$

We used a similar step in Example 15.3 to derive Laplace's equation. Whereas the Laplacian ∇^2 in the Dirichlet equation requires the second derivative of u, this expression only requires u to be *once* differentiable. In other words, we have expressed a second-order PDE in first-order language. Furthermore, this form of $\langle \cdot, \cdot\rangle_{\nabla^2}$ is symmetric and negative semidefinite, in the sense that

$$\langle u, u\rangle_{\nabla^2} = -\int_\Omega \|u(\vec{x})\|_2^2\,d\vec{x} \leq 0.$$

Our definition of weak PDE solutions above is far from formal. In particular, we were somewhat cavalier about the *space* of functions we should consider for u and v. Asking that $\langle u, v\rangle_{\nabla^2} = 0$ for all possible functions $v(\vec{x})$ is an unreasonable condition, since the space of all functions includes many degenerate functions that may not even be integrable. For the theoretical study of PDEs, it is usually sufficient to assume v is sufficiently smooth and has

small support. Even with this restriction, however, the space of functions is far too large to be discretized in any reasonable way.

The *finite elements method* (FEM), however, makes the construction above tractable by restricting functions to a finite basis. Suppose we wish to approximate u in a basis $\phi_1(\vec{x}), \ldots, \phi_k(\vec{x})$ by writing $u(\vec{x}) \approx \sum_{i=1}^{k} a_i \phi_i(\vec{x})$ for unknown coefficients $a_1, \ldots, a_k$. Since the actual solution $u(\vec{x})$ of the PDE is unlikely to be expressible in this form, we cannot expect $\langle \sum_i a_i \phi_i, v \rangle_{\nabla^2} = 0$ for all test functions $v(\vec{x})$. Hence, we not only approximate $u(\vec{x})$ but also restrict the class of test functions $v(\vec{x})$ to one in which we are more likely to be successful.

The best-known finite element approximation is the *Galerkin method*. In this method, we require that $\langle u, v \rangle_{\nabla^2} = 0$ for all test functions v that also can be written in the ϕ_i basis. By linearity of $\langle \cdot, \cdot \rangle_{\nabla^2}$, this method amounts to requiring that $\langle u, \phi_i \rangle_{\nabla^2} = 0$ for all $i \in \{1, \ldots, \Phi$. Expanding this relationship shows:

$$\begin{aligned} \langle u, \phi_i \rangle_{\nabla^2} &= \left\langle \sum_j a_j \phi_j, \phi_i \right\rangle_{\nabla^2} \text{ by our approximation of } u \\ &= \sum_j a_j \langle \phi_i, \phi_j \rangle_{\nabla^2} \text{ by linearity and symmetry of } \langle \cdot, \cdot \rangle_{\nabla^2}. \end{aligned}$$

In other words, we can recover the vector $\vec{a} \in \mathbb{R}^k$ of coefficients by solving the following linear system of equations:

$$\begin{pmatrix} \langle \phi_1, \phi_1 \rangle_{\nabla^2} & \langle \phi_1, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_1, \phi_k \rangle_{\nabla^2} \\ \langle \phi_2, \phi_1 \rangle_{\nabla^2} & \langle \phi_2, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_2, \phi_k \rangle_{\nabla^2} \\ \vdots & \vdots & \ddots & \vdots \\ \langle \phi_k, \phi_1 \rangle_{\nabla^2} & \langle \phi_k, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_k, \phi_k \rangle_{\nabla^2} \end{pmatrix} \vec{a} = \vec{0},$$

subject to the proper boundary conditions. For example, to impose nonzero Dirichlet boundary conditions, we can fix those values a_i corresponding to elements on the boundary $\partial\Omega$.

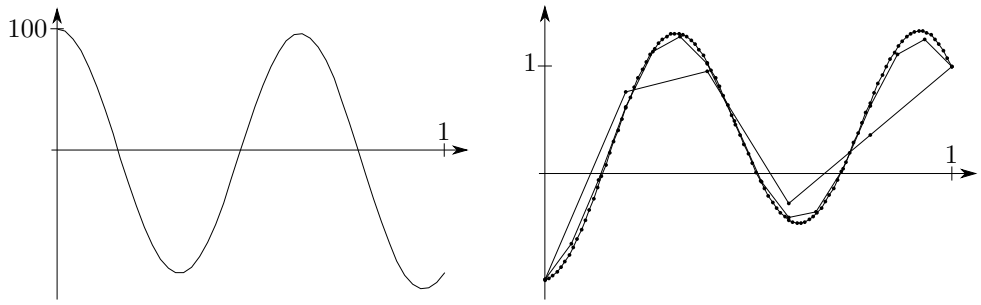
Approximating solutions to the Poisson equation $\nabla^2 u = w$ can be carried out in a similar fashion. If we write $w = \sum_i b_i \phi_i$, then Galerkin's method amounts to writing a slightly modified linear system of equations. The weak form of Poisson's equation has the same left-hand side but now has a nonzero right-hand side:

$$\int_\Omega v(\vec{x}) \nabla^2 u(\vec{x})\, d\vec{x} = \int_\Omega v(\vec{x}) w(\vec{x})\, d\vec{x}$$

for all test functions $v(\vec{x})$. To apply Galerkin's method in this case, we not only approximate $u(\vec{x}) = \sum_i a_i \phi_i(\vec{x})$ but also assume the right-hand side $w(\vec{x})$ can be written $w(\vec{x}) = \sum_i b_i \phi_i(\vec{x})$. Then, solving the weak Poisson equation in the ϕ_i basis amounts to solving:

$$\begin{pmatrix} \langle \phi_1, \phi_1 \rangle_{\nabla^2} & \langle \phi_1, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_1, \phi_k \rangle_{\nabla^2} \\ \langle \phi_2, \phi_1 \rangle_{\nabla^2} & \langle \phi_2, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_2, \phi_k \rangle_{\nabla^2} \\ \vdots & \vdots & \ddots & \vdots \\ \langle \phi_k, \phi_1 \rangle_{\nabla^2} & \langle \phi_k, \phi_2 \rangle_{\nabla^2} & \cdots & \langle \phi_k, \phi_k \rangle_{\nabla^2} \end{pmatrix} \vec{a} = \begin{pmatrix} \langle \phi_1, \phi_1 \rangle & \langle \phi_1, \phi_2 \rangle & \cdots & \langle \phi_1, \phi_k \rangle \\ \langle \phi_2, \phi_1 \rangle & \langle \phi_2, \phi_2 \rangle & \cdots & \langle \phi_2, \phi_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \phi_k, \phi_1 \rangle & \langle \phi_k, \phi_2 \rangle & \cdots & \langle \phi_k, \phi_k \rangle \end{pmatrix} \vec{b},$$

where $\langle f, g \rangle \equiv \int_\Omega f(\vec{x}) g(\vec{x})\, d\vec{x}$, the usual inner product of functions. The matrix next to $\vec{a}$ is known as the *stiffness matrix*, and the matrix next to $\vec{b}$ is known as the *mass matrix*. This is still a linear system of equations, since $\vec{b}$ is a fixed input to the Poisson equation.



$w(x)$ $u(x)$ (approx.)

FIGURE 15.17 Approximated piecewise linear solutions of $u''(x) = w(x)$ computed using finite elements as derived in Example 15.9; in these examples, we take $c = -1$, $d = 1$, and $k \in \{5, 15, 100\}$.

In practice, finite element methods such as Galerkin's method boil down to choosing appropriate spaces for approximation solutions u and test functions v. Once these spaces e.g. of ϕ_i's are chosen, the mass and stiffness matrices can be worked out offline, either in closed form or by using a quadrature method as explained in Chapter 13. These matrices are computable from the choice of basis functions. A few common choices are documented below:

- The most typical use case for FEM makes use of a triangulation of the domain Ω and takes the ϕ_i basis to be localized small neighborhoods of triangles. For example, for the Poisson equation it is sufficient to use piecewise-linear "hat" basis functions as discussed in §12.2.2 and illustrated in Figure 12.9. In this case, the mass and stiffness matrices are very sparse, because most of the basis functions ϕ_i have no overlap. Problem 15.2 works out the details of one such approach on the plane.
- Spectral methods use bases constructed out of cosine and sine, which have the advantage of being orthogonal with respect to $\langle \cdot, \cdot \rangle$; in particularly favorable cases, this orthogonality can make the mass or stiffness matrices diagonal. Furthermore, the fast Fourier transform and related algorithms accelerate computations in this case.
- Adaptive finite element methods analyze the output of a FEM solve to identify regions of Ω in which the solution has poor quality. Then, additional basis functions ϕ_i are added to refine the output in those regions.

Example 15.9 (Piecewise-linear FEM)**.** Suppose we wish to solve the Poisson equation $u''(x) = w(x)$ on the unit interval $x \in [0, 1]$ subject to Dirichlet boundary conditions $u(0) = c$ and $u(1) = d$. We will use the piecewise linear basis functions introduced in §12.1.3. In particular, define

$$\phi(x) \equiv \begin{cases} 1 + x & \text{when } x \in [-1, 0] \\ 1 - x & \text{when } x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

Then, for $k + 1$ basis elements we will define $\phi_i(kx - i) \equiv$ for $i \in \{0, \ldots, k\}$.

We work out the following integrals:

$$\int_{-1}^{1} \phi(x)^2\,dx = \int_{-1}^{0} (1+x)^2\,dx + \int_{0}^{1} (1-x)^2\,dx = \frac{2}{3}$$
$$\int_{-1}^{1} \phi(x)\phi(x-1)\,dx = \int_{0}^{1} x(1-x)\,dx = \frac{1}{6}$$

After applying change of coordinates, these integrals show:

$$\langle \phi_i, \phi_j \rangle = \frac{1}{k} \begin{cases} \frac{2}{3} & \text{when } i = j \\ \frac{1}{6} & \text{when } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the derivative $\phi'(x)$ satisfies:

$$\phi'(x) \equiv \begin{cases} 1 & \text{when } x \in [-1, 0] \\ -1 & \text{when } x \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

Hence, after change-of-variables we can write

$$\langle \phi_i, \phi_j \rangle_{d/dx} = -\langle \phi_i', \phi_j' \rangle = k \begin{cases} -2 & \text{when } i = j \\ 1 & \text{when } |i-j| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Up to the constant k, these values coincide with the divided difference second-derivative from §15.4.1.

We will apply the Galerkin method to discretize $u(x) \approx \sum_i a_i \phi_i(x)$. Assume we sample $b_i = w(i/k)$. Then, based on our integrals above, we should solve:

$$k \begin{pmatrix} 1/k & & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & & & 1/k \end{pmatrix} \vec{a} = \frac{1}{k} \begin{pmatrix} k & & & & & & \\ 1/6 & 2/3 & 1/6 & & & & \\ & 1/6 & 2/3 & 1/6 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1/6 & 2/3 & 1/6 & \\ & & & & & & k \end{pmatrix} \begin{pmatrix} c \\ b_1 \\ \vdots \\ b_{k-1} \\ d \end{pmatrix}$$

The first and last rows of this equation encode the boundary conditions, and the remaining rows come from the finite elements discretization. Figure 15.17 shows an example of this discretization in practice.

15.4.4 Finite Volumes

The *finite volume* method might be considered somewhere on the spectrum between finite elements and collocation. Like collocation, this method starts from the pointwise formulation of a PDE. Rather than asking that the PDE holds at a particular set of points in the domain Ω, however, finite volumes requires that the PDE is satisfied *on average* by integrating within the cells of a partition of Ω.

Suppose $\Gamma \subseteq \Omega$ is a region contained within the domain Ω and that we we once again wish to solve the Laplace equation $\nabla^2 u = 0$. A key tool for the finite volume method is the

divergence theorem, which states that the divergence of a vector field $\vec{v}(x)$ can be integrated over Γ two different ways:

$$\int_\Gamma \nabla \cdot \vec{v}(\vec{x})\, d\vec{x} = \int_{\partial\Gamma} \vec{v}(\vec{x}) \cdot \vec{n}(\vec{x})\, d\vec{x}.$$

Here, $\vec{n}$ is the normal to the boundary $\partial\Gamma$. In words, the divergence theorem states that integrating the divergence of a vector field in the interior of Γ is the same as integrating the normal component of Γ around the boundary. The latter is a measure of how much of $\vec{v}$ is "leaving" the region Γ, while the former sums how much $\vec{v}$ spreads out infinitesimally at each point inside Γ.

Suppose we solve the Poisson equation $\nabla^2 u = w$ in Ω. Then, within Γ we can write

$$\begin{aligned}
\int_\Gamma w(\vec{x})\, d\vec{x} &= \int_\Gamma \nabla^2 u(\vec{x})\, d\vec{x} \text{ since we solved the Poisson equation} \\
&= \int_\Gamma \nabla \cdot (\nabla u(\vec{x}))\, d\vec{x} \text{ since the Laplacian is the divergence of the gradient} \\
&= \int_{\partial\Gamma} \nabla u(\vec{x}) \cdot \vec{n}(\vec{x})\, d\vec{x} \text{ by the divergence theorem.}
\end{aligned}$$

Equating the first and last lines gives a condition for the Poisson equation averaged over Γ.

To derive a finite-volume approximation, again write $u(\vec{x}) \approx \sum_{i=1}^k a_i \phi_i(\vec{x})$ and now divide Ω into k regions $\Omega = \cup_{i=1}^k \Omega_i$. For each Ω_i we write:

$$\int_{\Omega_i} w(\vec{x})\, d\vec{x} = \int_{\partial\Omega_i} \nabla \left(\sum_{j=1}^k a_j \phi_j(\vec{x}) \right) \cdot \vec{n}(\vec{x})\, d\vec{x} = \sum_{j=1}^k a_j \left[\int_{\partial\Omega_i} \nabla \phi_j(\vec{x}) \cdot \vec{n}(\vec{x})\, d\vec{x} \right]$$

This is a linear system of equations for the a_i's. A typical discretization in this case might take the ϕ_i's to be piecewise-linear hat functions and the Ω_i's to be the Voronoi cells associated with the triangle centers (see §12.2.1).

15.4.5 Other Methods

Countless techniques exist for discretizing PDEs, and we have only scraped the surface of a few common methods in our discussion. Texts such as [73] are dedicated to developing the theoretical and practical aspects of these tools. Briefly, a few other notable methods for discretization include the following:

- *Domain decomposition* methods solve small versions of a PDE in different subregions of the domain Ω, iterating from one to the next until a solution to the global problem is reached. The subproblems can be made independent, in which case they are solvable via parallel processors. A single iteration of these methods can be used to approximate the solution of a PDE to precondition iterative solvers like conjugate gradients.
- The *boundary element* and *analytic element* methods solve certain PDEs using basis functions associated with points on the boundary $\partial\Omega$, reducing dependence on a triangulation or other discretization of the interior of Ω.
- *Mesh-free methods* simulate dynamical phenomena by tracking particles rather than meshing the domain. For example, the *smoothed-particle hydrodynamics* (SPH) technique in fluid simulation approximates a fluid as a collection of particles moving in space; particles can be added where additional detail is needed, and relatively few particles can be used to get realistic effects with limited computational capacity.

- *Level set methods*, used in image processing and fluid simulation, discretize PDEs governing the evolution and construction of curves and surfaces by representing those objects as level sets $\{\vec{x} \in \mathbb{R}^n : \psi(\vec{x}) = 0\}$. The unknown quantity then shifts from a shape to the level set function ψ.

15.5 SOLVING PARABOLIC AND HYPERBOLIC EQUATIONS

In the previous section, we mostly dealt with Poisson's equation, which is an elliptic PDE. Parabolic and hyperbolic equations generally introduce a *time* variable into the formulation, which also is differentiated but potentially to lower order. Discretizing time in the same fashion as space may not make sense for a given problem, since the two play fundamentally different roles in most physical phenomena. In this section, we consider options for discretizing this variable independently of the others.

15.5.1 Semidiscrete Methods

Semidiscrete methods apply the discretizations from §15.4 to the spatial domain but not time, leading to an ODE with a continuous time variable that can be solved using the methods of Chapter 14. This strategy is also known as the *method of lines*.

Example 15.10 (Semidiscrete heat equation)**.** Consider the heat equation in one variable, given by $u_t = u_{xx}$, where $u(t; x)$ represents the heat of a wire at position $x \in [0, 1]$ and time t. As boundary data, the user provides a function $u_0(x)$ such that $u(0; x) \equiv u_0(x)$; we also attach the boundary $x \in \{0, 1\}$ to a refrigerator and enforce Dirichlet conditions $u(t; 0) = u(t; 1) = 0$.

Suppose we discretize x using evenly-spaced samples but leave t as a continuous variable. If we use the finite differences technique from §15.4.1, this discretization results in functions $u_0(t), u_1(t), \ldots, u_n(t)$, where $u_i(t)$ represents the heat at position i as a function of time. Take L to be the corresponding second derivative matrix in the x samples with Dirichlet conditions. Then, the semidiscrete heat equation can be written $h^2\vec{u}'(t) = L\vec{u}(t)$, where $h = 1/n$ is the spacing between samples. This is an ODE for $\vec{u}(t)$ that could be time-stepped using backward Euler integration:

$$\vec{u}(t_{k+1}) \approx \left(I_{(n+1)\times(n+1)} - \frac{1}{h}L\right)^{-1} \vec{u}(t_k).$$

The previous example is an instance of a general pattern for parabolic equations. PDEs for *diffusive* phenomena like heat moving across a domain or chemicals moving through a membrane usually have one lower-order time variable and several spatial variables that are differentiated in an elliptic way. When we discretize the spatial variables using finite differences, finite elements, or another technique, the resulting semidiscrete formulation $\vec{u}' = A\vec{u}$ usually contains a negative definite matrix A. This makes the resulting ODE unconditionally stable.

As outlined in the previous chapter, we have many choices for solving the ODE after spatial discretization. If time steps are small, explicit methods may be acceptable. Implicit solvers, however, often are applied to solving parabolic PDEs; diffusive behavior of implicit Euler agrees behaviorally with diffusion from the heat equation and may be acceptable even with fairly large time steps. Hyperbolic PDEs, on the other hand, may require implicit steps for stability, but advanced integrators such as "symplectic integrators" can prevent oversmoothing of non-diffusive phenomena.

When A does not change with time, one contrasting approach is to write solutions of semidiscrete systems $\vec{u}' = A\vec{u}$ in terms of eigenvectors of A. Suppose $\vec{v}_1, \ldots, \vec{v}_n$ are eigenvectors of A with eigenvalues $\lambda_1, \ldots, \lambda_n$ and that we know $\vec{u}(0) = c_1\vec{v}_1 + \cdots + c_n\vec{v}_n$. Then, as we showed in §5.1.2, the solution of $\vec{u}' = A\vec{u}$ is given by:

$$\vec{u}(t) = \sum_i c_i e^{\lambda_i t}\vec{v}_i$$

The eigenvectors and eigenvalues of A may have physical interpretations in the case of a semidiscrete PDE. Most commonly, the eigenvalues the Laplacian ∇^2 on a domain Ω correspond to resonant frequencies of a domain, that is, the frequencies that sound when hitting the domain with a hammer. Thus, the eigenvectors provide closed-form "low-frequency approximations" of solutions to common PDEs after truncating the sum above over i.

15.5.2 Fully Discrete Methods

Rather than discretizing time and then space, we might treat the space and time variables more democratically and discretize them both simultaneously. This one-shot discretization is in some sense a more direct application of the methods we considered in §15.4, just by including t as a dimension in the domain Ω under consideration. Because we now multiply the number of variables needed to represent Ω by the number of time steps, the resulting linear systems of equations can be large if dependence between time steps has global reach.

Example 15.11 (Fully-discrete heat diffusion, [54])**.** Consider the heat equation $u_t = u_{xx}$. Discretizing x and t simultaneously via finite differences yields a matrix of u values, which we can index u_i^j, representing the heat at position i and time j. Take Δx and Δt to be the spacing of x and t in the grid, resp. If we wish to step from time j to time $j+1$, choosing where to evaluate the different derivatives brings different discretization schemes.

For example, evaluating the x derivative at time j produces an *explicit* formula:

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} = \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{(\Delta x)^2}.$$

Isolating u_i^{j+1} gives a formula for obtaining u at time $j+1$ without a linear solve.

Alternatively, we can evaluate the x derivative at time $j+1$ for an *implicit* heat equation integrator:

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} = \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{(\Delta x)^2}.$$

This integrator is unconditionally stable but requires a linear solve to obtain the u values at time $j+1$ from those at time j.

The implicit and explicit heat equation integrators inherit their accuracy from the quality of the finite difference formulas, and hence—stability aside—both are first-order accurate in time and second-order accurate in space. To improve the accuracy of the time discretization, we can use the *Crank-Nicolson* method, which applies a trapezoidal time integrator:

$$\frac{u_i^{j+1} - u_i^j}{\Delta t} = \frac{1}{2}\left[\frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{(\Delta x)^2} + \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{(\Delta x)^2}\right].$$

This method inherits the unconditional stability of the trapezoidal integrator and is second-order accurate in time and space. Despite this stability, however, as explained in §14.3.3 taking time steps that are too large can produce unrealistic oscillatory behavior.

In the end, even semidiscrete methods can be considered fully discrete in that the time-stepping ODE method still discretizes the t variable; the difference between semidiscrete and fully-discrete is mostly for classification of how methods were derived. One advantage of semidiscrete techniques, however, is that they can adjust the time step for t depending on the current iterate, e.g. if objects are moving quickly in a physical simulation it might make sense to take more dense time steps and resolve this motion. Some methods also adjust the discretization of the domain of x values in case more resolution is needed near local discontinuities such as shock waves.

15.6 NUMERICAL CONSIDERATIONS

We have considered several options for discretizing PDEs. As with choosing time integrators for ODEs, the trade-offs between these options are intricate, representing different compromises between computational efficiency, numerical quality, stability, and so on. The situation is even more complex for PDEs, which include ODEs as a special case. We conclude our consideration of numerical methods for PDE by outlining a few considerations when choosing a PDE discretization.

15.6.1 Consistency, Convergence, and Stability

A key consideration when choosing ODE integrators was *stability*, which guaranteed that errors in specifying initial conditions would not be amplified over time. Stability remains a consideration in PDE integration, but it also can interact with other key properties:

- A method is *convergent* if solutions to the discretized problem converge to the theoretical solution of the PDE as spacing between discrete samples approaches zero.
- A method is *consistent* if the accompanying discretization of the differential operators better approximates the derivatives taken in the PDE as spacing approaches zero.

For finite differencing schemes, the Lax-Richtmyer Equivalence Theorem states that if a linear problem is well-posed, *consistency* and *stability* together are necessary and sufficient for *convergence* [74]. Consistency and stability tend to be easier to check than convergence. Consistency arguments usually come from Taylor series. A number of well-established methods establish stability or lack thereof; for example, the well-known *CFL condition* states that the ratio of time spacing to spatial spacing of samples should exceed the speed at which waves propagate in the case of hyperbolic PDE [27]. Even more caution must be taken when simulating advective phenomena and PDEs that can develop fronts and shocks; specialized *upwinding* schemes attempt to detect the formation of these features to ensure that they move in the right direction and at the proper speed.

Even when a time variable is not involved, some care must be taken to ensure that a PDE approximation scheme reduces error as sampling becomes more dense. For example, in elliptic PDE, convergence of finite elements methods depends on the choice of basis functions, which must be sufficiently smooth to represent the theoretical solution and must span the function space in the limit [14].

The subtleties of consistency, convergence, and stability underlie much of the theory in numerical PDE, and the importance of these concepts cannot be overstated. After all, without a convergence guarantee, the output of a numerical PDE solver cannot be trusted. Hence, standard PDE integration packages often incorporate checks for assorted stability conditions or degenerate behavior to help guide clients of the software whose expertise is in modeling rather than numerics.

15.6.2 Linear Solvers for PDE

The matrices resulting from PDE discretizations have many favorable properties that make them ideal applications of the methods we have considered in previous chapters. For instance, as motivated in §15.3.1, elliptic PDEs are closely-related to positive definite matrices, and typical discretizations typically require solution of a positive definite linear system. The same derivative operators appear in parabolic PDEs, which hence have well-posed semidiscretizations. Hence, methods like Cholesky decomposition and conjugate gradients typically are applied to these problems. Furthermore, derivative matrices tend to be highly sparse, inducing additional memory and time savings. Any reasonable implementation of a PDE solver should include these sorts of efficiency optimizations, which make them scalable to much larger problems.

Example 15.12 (Elliptic operators as matrices)**.** Consider the one-dimensional second derivative matrix L with Dirichlet boundary conditions from §15.4.1. L is sparse and negative definite. To show the latter property, we can write $L = -D^\top D$ for the matrix $D \in \mathbb{R}^{(n+1)\times n}$ given by:

$$D = \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & \end{pmatrix}$$

This matrix is a finite-differenced *first* derivative operator, so this observation parallels the fact that $d^2y/dx^2 = d/dx(dy/dx)$. For any $\vec{x} \in \mathbb{R}^n$, $\vec{x}^\top L\vec{x} = -\vec{x}^\top D^\top D\vec{x} = -\|D\vec{x}\|_2^2 \leq 0$, showing L is negative semidefinite. Furthermore, $D\vec{x} = 0$ exactly when $\vec{x} = 0$, completing the proof that L is negative definite.

Example 15.13 (Stiffness matrix is positive semidefinite)**.** Regardless of the basis $\phi_1, \ldots, \phi_k$, the stiffness matrix that comes from discretizing Poisson's equation via finite elements (see §15.4.3) is negative semidefinite. If the define M_{∇^2} to be the stiffness matrix, then for $\vec{a} \in \mathbb{R}^k$ we can write:

$$\begin{aligned}
\vec{a}^\top M_{\nabla^2}\vec{a} &= \sum_{ij} a_i a_j \langle \phi_i, \phi_j \rangle_{\nabla^2} \text{ by definition of } M_{\nabla^2} \\
&= \left\langle \sum_i a_i\phi_i, \sum_j a_j\phi_j \right\rangle_{\nabla^2} \text{ by bilinearity of } \langle\cdot,\cdot\rangle_{\nabla^2} \\
&= \langle \psi, \psi \rangle_{\nabla^2} \text{ if we define } \psi \equiv \sum_i a_i\phi_i \\
&= -\int_\Omega \|\psi(\vec{x})\|_2^2 \, d\vec{x} \text{ by definition of } \langle\cdot,\cdot\rangle_{\nabla^2} \\
&\leq 0
\end{aligned}$$

15.7 EXERCISES

15.1 ("Shooting method," [54]) The *two-point boundary value problem* inherits some structure from ODE and PDE problems alike. In this problem, we wish to solve the ODE

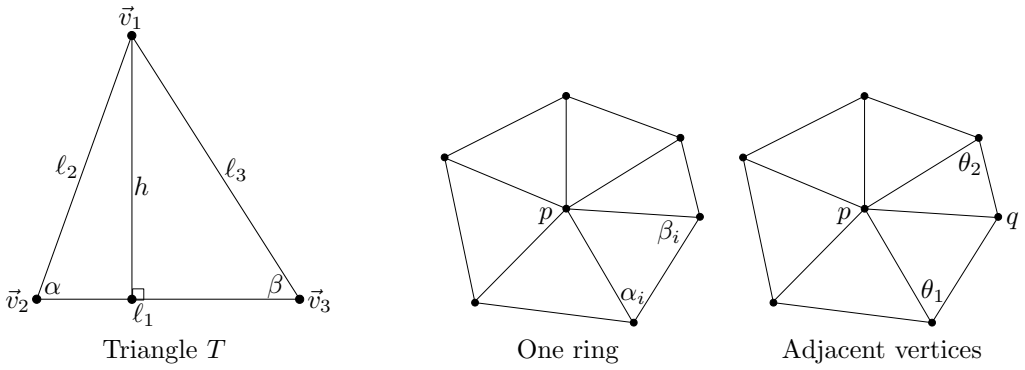


FIGURE 15.18 Notation for problem 15.2.

$\vec{y}' = F[\vec{y}]$ for a function $\vec{y}(t) : [0,1] \to \mathbb{R}^n$. Rather than specifying initial conditions, however, we specify some relationship $g(\vec{y}(0), \vec{y}(1)) = \vec{0}$.

(a) Give a nontrivial example of a two-point boundary value problem that does not admit a solution.

(b) Assume we have checked the conditions of an existence/uniqueness theorem, so given $\vec{y}_0 = \vec{y}(0)$ we can generate $\vec{y}(t)$ for all $t > 0$ satisfying $\vec{y}'(t) = F[\vec{y}(t)]$. Denote $\vec{y}(t; \vec{y}_0) : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}$ as the function returning $\vec{y}$ at time t given $\vec{y}(0) = \vec{y}_0$. In this notation, pose the two-point boundary value problem as a root-finding problem.

(c) Use the ODE integration methods from Chapter 14 to propose a computationally feasible root-finding problem for approximating a solution $\vec{y}(t)$ of the two-point boundary value problem.

(d) As discussed in Chapter 7, most root-finding algorithms require the Jacobian of the objective function. Suggest a technique for finding the Jacobian of your objective from 15.1c.

15.2 In this problem, we use first-order finite elements to derive the famous *cotangent Laplacian* formula used in geometry processing. Refer to Figure 15.18 for notation.

(a) Suppose we construct a planar triangle T with vertices $\vec{v}_1, \vec{v}_2, \vec{v}_3 \in \mathbb{R}^2$ in counterclockwise order. Take $f_1(\vec{x})$ to be the affine *hat function* $f_1(\vec{x}) \equiv c + \vec{d} \cdot \vec{x}$ satisfying $f_1(\vec{v}_1) = 1$, $f_1(\vec{v}_2) = 0$, and $f_1(\vec{v}_3) = 0$. Show that ∇f_1 is a constant vector satisfying:

$$\begin{aligned}\nabla f_1 \cdot (\vec{v}_1 - \vec{v}_2) &= 1\\ \nabla f_1 \cdot (\vec{v}_1 - \vec{v}_3) &= 1\\ \nabla f_1 \cdot (\vec{v}_2 - \vec{v}_3) &= 0\end{aligned}$$

Conclude that ∇f_1 is perpendicular to the edge from $\vec{v}_2$ to $\vec{v}_3$.

(b) Show that $\|\nabla f_1\|_2 = \frac{1}{h}$, where h is the height of the triangle as marked in Figure 15.18 (left).
Hint: Start by showing $\nabla f_1 \cdot (\vec{v}_1 - \vec{v}_3) = \|\nabla f_1\|_2 \ell_3 \cos\left(\frac{\pi}{2} - \beta\right)$.

(c) Integrate over the triangle T to show

$$\int_T \|\nabla f_1\|_2^2 \, dA = \frac{1}{2}(\cot\alpha + \cot\beta).$$

Hint: Since ∇f_1 is a constant vector, the integral equals $\|\nabla f_1\|_2^2 A$, where A is the area of T. From basic geometry, we know $A = \frac{1}{2}\ell_1 h$.

(d) Define $\theta \equiv= \pi - \alpha - \beta$, and take f_2 and f_3 to be the hat functions associated with $\vec{v}_2$ and $\vec{v}_3$, resp. Show that

$$\int_T \nabla f_2 \cdot \nabla f_3 \, dA = -\frac{1}{2}\cot\theta.$$

(e) Now, consider a vertex p of a triangle mesh (Figure 15.18, middle), and define $f_p : \mathbb{R}^2 \to [0,1]$ to be the *piecewise linear hat function* associated with p (see §12.2.2 and Figure 12.9). That is, restricted to any triangle adjacent to p, the function f_p behaves as constructed in 15.2a; $f_p \equiv 0$ outside the triangles adjacent to p. Based on the results you already have constructed, show:

$$\int_{\mathbb{R}^2} \|\nabla f_p\|_2^2 \, dA = \frac{1}{2}\sum_i (\cot\alpha_i + \cot\beta_i),$$

where $\{\alpha_i\}$ and $\{\beta_i\}$ are the angles opposite p in its neighboring triangles.

(f) Now, suppose p and q are adjacent vertices on the same mesh, and define θ_1 and θ_2 as shown in Figure 15.18 (right). Show

$$\int_{\mathbb{R}^2} \nabla f_p \cdot \nabla f_q \, dA = -\frac{1}{2}(\cot\theta_1 + \cot\theta_2).$$

(g) Conclude that in the basis of hat functions on a triangle mesh, the stiffness matrix for the Poisson equation has the following form:

$$L_{ij} \equiv -\frac{1}{2}\begin{cases} \sum_{i\sim j}(\cot\alpha_j + \cot\beta_j) & \text{if } i = j \\ -(\cot\alpha_j + \cot\beta_j) & \text{if } i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

Here, $i \sim j$ denotes that vertices i and j are adjacent.

(h) Write a formula for the entries of the corresponding mass matrix, whose entries are

$$\int_{\mathbb{R}^2} f_p f_q \, dA.$$

Hint: This matrix can be written completely in terms of triangle areas. Divide into cases: (1) $p = q$, (2) p and q are adjacent vertices, and (3) p and q are not adjacent.

15.3 Suppose we wish to approximate Laplacian eigenfunctions $f(\vec{x})$, satisfying $\nabla^2 f = \lambda f$. Show that discretizing such a problem using FEM results in a generalized eigenvalue problem $A\vec{x} = \lambda B\vec{x}$.

15.4 Propose a semidiscrete form for the one-dimensional wave equation $u_{tt} = u_{xx}$, similar to the construction in Example 15.10. Is the resulting ODE well-posed (§14.2.3)?

15.5 *Graph-based semi-supervised learning* algorithms attempt to predict a quantity or label associated with the nodes of a graph given labels on a few of its vertices. For instance, under the (dubious) assumption that friends are likely to have similar incomes, it could be used to predict the annual incomes of all members of a social network given the incomes of a few of its members. We will focus on a variation of the method proposed in [126].

(a) Take $G = (V, E)$ to be a connected graph, and define $f_0 : V_0 \to \mathbb{R}$ to be a set of scalar-valued labels associated with the nodes of a subset $V_0 \subseteq V$. The *Dirichlet energy* of a full assignment of labels $f : V \to \mathbb{R}$ is given by

$$E[f] \equiv \sum_{(v_1,v_2)\in E} (f(v_2) - f(v_1))^2.$$

Explain why $E[f]$ can be minimized over f satisfying $f(v_0) = f_0(v_0)$ for all $v_0 \in V_0$ using a linear solve.

(b) Explain the connection between the linear system from 15.5a and the 3×3 Laplacian stencil from §15.4.1.

(c) Suppose f the result of the optimization from 15.5a. Prove the *discrete maximum principle*:

$$\max_{v\in V} f(v) = \max_{v_0\in V_0} f_0(v_0).$$

Relate this result to a physical interpretation of Laplace's equation.

15.6 Give an example where discretization of the Poisson equation via finite differences and via collocation lead to the same system of equations.

15.7 ("Von Neumann stability analysis," based on notes by D. Levy) Suppose we wish to approximate solutions to the PDE $u_t = au_x$ for some fixed $a \in \mathbb{R}$. We will use initial conditions $u(x, 0) = f(x)$ for some $f \in C^\infty([0, 2\pi])$ and periodic boundary conditions $u(0, t) = u(2\pi, t)$.

(a) What is the order of this PDE? Give conditions on a for it to be elliptic, hyperbolic, or parabolic.

(b) Show that the PDE is solved by $u(x, t) = f(x + at)$.

(c) The Fourier transform of $u(x, t)$ in x is

$$[\mathcal{F}_x u](\omega, t) \equiv \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} u(x, t)e^{-i\omega x}\, dx,$$

where $i = \sqrt{-1}$ (see problem 3.13). It measures the frequency content of $u(\cdot, t)$. Define $v(x, t) \equiv u(x + \Delta x, t)$. Show that $[\mathcal{F}_x v](\omega, t) = e^{i\omega\Delta x}[\mathcal{F}_x u](\omega, t)$.

(d) Suppose we use a forward Euler discretization:

$$\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = a\frac{u(x + \Delta x, t) - u(x - \Delta x, t)}{2\Delta x}.$$

Show that this discretization satisfies

$$[\mathcal{F}_x u](\omega, t + \Delta t) = \left(1 + \frac{ai\Delta t}{\Delta x} \sin(\omega\Delta x)\right) [\mathcal{F}_x u](\omega, t).$$

(e) Define the *amplification factor*

$$\hat{Q} \equiv 1 + \frac{ai\Delta t}{\Delta x}\sin(\omega\Delta x).$$

Show that $|\hat{Q}| > 1$. This shows that the discretization amplifies frequency content over time and is *unconditionally unstable*.

(f) Carry out a similar analysis for the alternative discretization

$$u(x,t+\Delta t) = \frac{1}{2}\left(u(x-\Delta x,t) + u(x+\Delta x,t)\right) + \frac{a\Delta t}{2\Delta x}\left[u(x+\Delta x,t) - u(x-\Delta x,t)\right].$$

Derive an upper bound on the ratio $\Delta t/\Delta x$ for this discretization to be stable.

15.8 ("Fast marching," [17]) Nonlinear PDEs require specialized treatment. One nonlinear PDE relevant to computer graphics and medical imaging is the eikonal equation $\|d\|_2 = 1$ considered in §15.5. Here, we outline some aspects of the *fast marching* method for solving this equation on a triangulated domain $\Omega \subset \mathbb{R}^2$ (see Figure 12.9).

(a) We might approximate solutions of the eikonal equation as shortest-path distances along the edges of the triangulation. Provide triangulations of the unit square $[0,1] \times [0,1]$ with arbitrarily small triangle edge lengths and areas for which this approximation gives distance 2 rather than $\sqrt{2}$ from $(0,0)$ to $(1,1)$. Can the approximation be consider convergent?

(b) Suppose we approximate $d(\vec{x})$ with a linear function $d(\vec{x}) \approx \vec{n}^\top\vec{x} + p$, where $\|\vec{n}\|_2 = 1$ by the eikonal equation. Given $d_1 = d(\vec{x}_1)$ and $d_2 = d(\vec{x}_2)$, show that p can be recovered by solving a quadratic equation and provide a geometric interpretation of the two roots.

(c) What geometric assumption does the approximation in 15.8b make about the shape of the level sets $\{\vec{x} \in \mathbb{R}^2 : d(\vec{x}) = c\}$? Does this approximation make sense when d is large or small? See [86] for a contrasting circular approximation.

(d) Extend Dijkstra's algorithm for graph-based shortest paths to triangulated shapes using the approximation in 15.8b. What can go wrong with this approach?

Hint: Dijkstra's algorithm starts at the center vertex and builds the shortest path in breadth-first fashion. Change the update to use 15.8b, and consider when the approximation will make distances decrease unnaturally.

15.9 Constructing *higher-order elements* can be necessary for solving certain differential equations.

(a) Show that the parameters $a_0, \ldots, a_5$ of a function $f(x,y) = a_0 + a_1x + a_2y + a_3x^2 + a_4y^2 + a_5xy$ are uniquely determined by its values on the three vertices and three edge midpoints of a triangle.

(b) Show that if (x,y) is on an edge of the triangle, then $f(x,y)$ can be computed knowing only the values of f at the endpoints and midpoint of that edge.

(c) Use these facts to construct a basis of continuous, piecewise-quadratic functions on a triangle mesh, and explain why it may be useful for solving higher-order PDEs.

15.10 For matrices $A, B \in \mathbb{R}^{n\times n}$, the Lie-Trotter-Kato formula states

$$e^{A+B} = \lim_{n\to\infty} (e^{A/n} e^{B/n})^n,$$

where e^M denotes the matrix exponential of $M \in \mathbb{R}^{n\times n}$ (see §14.3.5).

Suppose we wish to solve a PDE $u_t = \mathcal{L}u$, where $\mathcal{L}$ is some differential operator that admits a *splitting* $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$. How can the Lie-Trotter-Kato formula be applied to designing PDE time-stepping machinery in this case?

Note: Such splittings are useful for breaking up integrators for complex PDEs like the Navier-Stokes equations into simpler steps.

Bibliography

[1] S. Ahn, U. J. Choi, and A. G. Ramm. A scheme for stable numerical differentiation. *Journal of Computational and Applied Mathematics*, 186(2):325–334, 2006.

[2] E. Anderson, Z. Bai, and J. Dongarra. Generalized QR factorization and its applications. *Linear Algebra and its Applications*, 162–164(0):243–271, 1992.

[3] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[4] S. Axler. Down with determinants! *American Mathematical Monthly*, 102:139–154, 1995.

[5] D. Baraff, A. Witkin, and M. Kass. Untangling cloth. *ACM Transactions on Graphics*, 22(3):862–870, July 2003.

[6] J. Barbič and Y. Zhao. Real-time large-deformation substructuring. *ACM Transactions on Graphics*, 30(4):91:1–91:8, July 2011.

[7] R. Barrett, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial and Applied Mathematics, 1994.

[8] M. Bartholomew-Biggs, S. Brown, B. Christianson, and L. Dixon. Automatic differentiation of algorithms. *Journal of Computational and Applied Mathematics*, 124(12):171–190, 2000.

[9] H. Bauschke and J. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[10] H. H. Bauschke and Y. Lucet. What is a Fenchel conjugate? *Notices of the American Mathematical Society*, 59(1), 2012.

[11] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, Mar. 2009.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, Jan. 2011.

[13] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[14] S. Brenner and R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer, 2008.

[15] R. Brent. *Algorithms for Minimization Without Derivatives*. Dover Books on Mathematics. Dover, 2013.

[16] J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.

[17] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Monographs in Computer Science. Springer, 2008.

[18] S. Bubeck. Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980*, 2014.

[19] C. Budd. Advanced numerical methods (MA50174): Assignment 3, initial value ordinary differential equations. University Lecture, 2006.

[20] R. Burden and J. Faires. *Numerical Analysis*. Cengage Learning, 2010.

[21] W. Cheney and A. A. Goldstein. Proximity maps for convex sets. *Proceedings of the American Mathematical Society*, 10(3):448–450, 1959.

[22] M. Chuang and M. Kazhdan. Interactive and anisotropic geometry processing using the screened Poisson equation. *ACM Transactions on Graphics*, 30(4):57:1–57:10, July 2011.

[23] C. Clenshaw and A. Curtis. A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2(1):197–205, 1960.

[24] A. Colorni, M. Dorigo, and V. Maniezzo. Distributed optimization by ant colonies. In *Proceedings of the European Conference on Artificial Life*, pages 134–142, 1991.

[25] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.

[26] P. G. Constantine and D. F. Gleich. Tall and skinny QR factorizations in MapReduce architectures. In *Proceedings of the International Workshop on MapReduce and Its Applications*, pages 43–50, 2011.

[27] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74, 1928.

[28] Y. H. Dai and Y. Yuan. A nonlinear conjugate gradient method with a strong global convergence property. *SIAM Journal on Optimization*, 10(1):177–182, May 1999.

[29] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

[30] M. de Berg. *Computational Geometry: Algorithms and Applications*. Springer, 2000.

[31] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011.

[32] S. T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[33] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. In *Micro Machine and Human Science*, pages 39–43, Oct 1995.

[34] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer, 2010.

[35] M. A. Epelman. Continuous optimization methods (IOE 511): Rate of convergence of the steepest descent algorithm. University Lecture, 2007.

[36] E. Fehlberg. *Low-order classical Runge-Kutta formulas with stepsize control and their application to some heat transfer problems.* NASA technical report. National Aeronautics and Space Administration, 1969.

[37] R. Fletcher. Conjugate gradient methods for indefinite systems. In G. A. Watson, editor, *Numerical Analysis*, volume 506 of *Lecture Notes in Mathematics*, pages 73–89. Springer, 1976.

[38] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2):149–154, 1964.

[39] D. C.-L. Fong and M. Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, Oct. 2011.

[40] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956.

[41] R. W. Freund and N. M. Nachtigal. QMR: A quasi-minimal residual method for non-hermitian linear systems. *Numerische Mathematik*, 60(1):315–339, 1991.

[42] C. Führer. Numerical methods in mechanics (FMN 081): Homotopy method. University Lecture, 2006.

[43] M. Geradin and D. Rixen. *Mechanical Vibrations: Theory and Application to Structural Dynamics.* Wiley, 1997.

[44] W. Givens. Computation of plane unitary rotations transforming a general matrix to triangular form. *Journal of the Society for Industrial and Applied Mathematics*, 6(1):26–50, 1958.

[45] D. Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, Mar. 1991.

[46] G. Golub and C. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2012.

[47] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1.

[48] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer, 2008.

[49] E. Grinspun and M. Wardetzky. Discrete differential geometry: An applied introduction. In *SIGGRAPH Asia Courses*, 2008.

[50] C. W. Groetsch. Lanczos' generalized derivative. *American Mathematical Monthly*, 105(4):320–326, 1998.

[51] L. Guibas, D. Salesin, and J. Stolfi. Epsilon geometry: Building robust algorithms from imprecise computations. In *Proceedings of the Symposium on Computational Geometry*, pages 208–217, 1989.

[52] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Applied Mathematical Sciences. Springer, 1993.

[53] G. Hairer. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, 2010.

[54] M. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill, 2005.

[55] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, Dec. 1952.

[56] D. J. Higham and L. N. Trefethen. Stiffness of ODEs. *BIT Numerical Mathematics*, 33(2):285–303, 1993.

[57] N. J. Higham. Computing the polar decomposition with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, Oct. 1986.

[58] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, Aug. 2002.

[59] M. Hirsch, S. Smale, and R. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 3rd edition, 2012.

[60] A. S. Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM*, 5(4):339–342, Oct. 1958.

[61] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, volume 28, pages 427–435, 2013.

[62] D. L. James and C. D. Twigg. Skinning mesh animations. *ACM Transactions on Graphics*, 24(3):399–407, July 2005.

[63] F. John. The ultrahyperbolic differential equation with four independent variables. *Duke Mathematical Journal*, 4(2):300–322, 6 1938.

[64] W. Kahan. Pracniques: Further remarks on reducing truncation errors. *Communications of the ACM*, 8(1):47–48, Jan. 1965.

[65] J. T. Kajiya. The rendering equation. In *Proceedings of SIGGRAPH*, volume 20, pages 143–150, 1986.

[66] Q. Ke and T. Kanade. Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2005.

[67] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE Conference on Neural Networks*, volume 4, pages 1942–1948, Nov 1995.

[68] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[69] K. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*. Lecture Notes in Mathematics. Springer, 1985.

[70] A. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. In *Numerical Treatment of Eigenvalue Problems*, volume 5, pages 143–154. Springer, 1991.

[71] A. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[72] C. Lanczos. *Applied Analysis*. Dover Books on Mathematics. Dover Publications, 1988.

[73] S. Larsson and V. Thomee. *Partial Differential Equations with Numerical Methods*. Texts in Applied Mathematics. Springer, 2008.

[74] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on Pure and Applied Mathematics*, 9(2):267–293, 1956.

[75] R. B. Lehoucq and D. C. Sorensen. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, Oct. 1996.

[76] M. Leordeanu and M. Hebert. Smoothing-based optimization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, June 2008.

[77] K. Levenberg. A method for the solution of certain non-linear problems in least-squares. *Quarterly of Applied Mathematics*, 2(2):164–168, July 1944.

[78] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(13):193–228, 1998.

[79] D. Luenberger and Y. Ye. *Linear and Nonlinear Programming*. International Series in Operations Research & Management Science. Springer, 2008.

[80] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[81] J. McCann and N. S. Pollard. Real-time gradient-domain painting. *ACM Transactions on Graphics*, 27(3):93:1–93:7, Aug. 2008.

[82] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1998.

[83] Y. Nesterov and I. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

[84] J. Niesen and W. M. Wright. Algorithm 919: A Krylov subspace algorithm for evaluating the φ-functions appearing in exponential integrators. *ACM Transactions on Mathematical Software*, 38(3):22:1–22:19, Apr. 2012.

[85] J. Nocedal and S. Wright. *Numerical Optimization*. Series in Operations Research and Financial Engineering. Springer, 2006.

[86] M. Novotni and R. Klein. Computing geodesic distances on triangular meshes. In *Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, Feb. 2002.

[87] J. M. Ortega and H. F. Kaiser. The LLT and QR methods for symmetric tridiagonal matrices. *The Computer Journal*, 6(1):99–101, 1963.

[88] C. Paige and M. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[89] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, Mar. 1982.

[90] T. Papadopoulo and M. I. A. Lourakis. Estimating the Jacobian of the singular value decomposition: Theory and applications. In *Proceedings of the European Conference on Computer Vision*, pages 554–570. Springer, 2000.

[91] S. Paris, P. Kornprobst, and J. Tumblin. *Bilateral Filtering: Theory and Applications*. Foundations and Trends in Computer Graphics and Vision. Now Publishers, 2009.

[92] S. Paris, P. Kornprobst, J. Tumblin, and F. Durand. A gentle introduction to bilateral filtering and its applications. In *ACM SIGGRAPH 2007 Courses*, 2007.

[93] B. N. Parlett and J. Poole, W. G. A geometric theory for the QR, LU and power iterations. *SIAM Journal on Numerical Analysis*, 10(2):389–412, 1973.

[94] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012.

[95] E. Polak and G. Ribière. Note sur la convergence de methodes de directions conjuguees. *Modelisation Mathematique et Analyse Numerique*, 3(R1):35–43, 1969.

[96] W. Press. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[97] L. Ramshaw. *Blossoming: A Connect-the-Dots Approach to Splines*. Number 19 in SRC Reports. Digital Equipment Corporation, 1987.

[98] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, 2005.

[99] R. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[100] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2nd edition, 2003.

[101] Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, July 1986.

[102] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[103] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the ACM National Conference*, pages 517–524, 1968.

[104] J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, Carnegie Mellon University, 1994.

[105] J. Shi and J. Malik. Normalized cuts and image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, Aug 2000.

[106] K. Shoemake and T. Duff. Matrix animation and polar decomposition. In *Proceedings of the Conference on Graphics Interface*, pages 258–264, 1992.

[107] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayński. *Minimization Methods for Non-differentiable Functions*. Springer, 1985.

[108] M. Slawski and M. Hein. Sparse recovery by thresholded non-negative least squares. In *Advances in Neural Information Processing Systems*, pages 1926–1934, 2011.

[109] P. Sonneveld. CGS: A fast Lanczos-type solver for nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 10(1):36–52, 1989.

[110] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Symposium on Geometry Processing*, pages 109–116. Eurographics Association, 2007.

[111] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Texts in Applied Mathematics. Springer, 2002.

[112] L. H. Thomas. Elliptic problems in linear differential equations over a network. Technical report, Columbia University, 1949.

[113] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[114] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the International Conference on Computer Vision*, pages 839–846, 1998.

[115] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Symposium on Discrete Algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

[116] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, Jan. 1991.

[117] W. T. Tutte. How to draw a graph. *Proceedings of the London Mathematical Society*, 13(1):743–767, 1963.

[118] H. Uzawa and K. Arrow. *Iterative Methods for Concave Programming*. Cambridge University Press, 1989.

[119] J. van de Weijer and R. van den Boomgaard. Local mode filtering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 428–433, 2001.

[120] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, Mar. 1992.

[121] S. Wang and L. Liao. Decomposition method with a variable parameter for a class of monotone variational inequality problems. *Journal of Optimization Theory and Applications*, 109(2):415–429, 2001.

[122] M. Wardetzky, S. Mathur, F. Kälberer, and E. Grinspun. Discrete Laplace operators: No free lunch. In *Proceedings of the Symposium on Geometry Processing*, pages 33–37, 2007.

[123] O. Weber, M. Ben-Chen, and C. Gotsman. Complex barycentric coordinates with applications to planar shape deformation. *Computer Graphics Forum*, 28(2), 2009.

[124] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, Oct. 2006.

[125] J. H. Wilkinson. *The perfidious polynomial*. Mathematical Association of America, 1984.

[126] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning*, volume 3, pages 912–919, 2003.