

Undergraduate Texts in Mathematics

UTM

Peter J. Olver · Chehrzad Shakiban

# Applied Linear Algebra

*Second Edition*



Springer

# Undergraduate Texts in Mathematics

# Undergraduate Texts in Mathematics

---

## Series Editors:

Sheldon Axler  
*San Francisco State University, San Francisco, CA, USA*

Kenneth Ribet  
*University of California, Berkeley, CA, USA*

## Advisory Board:

Colin Adams, *Williams College*  
David A. Cox, *Amherst College*  
L. Craig Evans, *University of California, Berkeley*  
Pamela Gorkin, *Bucknell University*  
Roger E. Howe, *Yale University*  
Michael Orrison, *Harvey Mudd College*  
Lisette G. de Pillis, *Harvey Mudd College*  
Jill Pipher, *Brown University*  
Fadil Santosa, *University of Minnesota*

**Undergraduate Texts in Mathematics** are generally aimed at third- and fourth-year undergraduate mathematics students at North American universities. These texts strive to provide students and teachers with new perspectives and novel approaches. The books include motivation that guides the reader to an appreciation of interrelations among different aspects of the subject. They feature examples that illustrate key concepts as well as exercises that strengthen understanding.

More information about this series at <http://www.springer.com/series/666>

Peter J. Olver • Chehrzad Shakiban

# Applied Linear Algebra

Second Edition



Springer

Peter J. Olver  
School of Mathematics  
University of Minnesota  
Minneapolis, MN  
USA

Chehrzad Shakiban  
Department of Mathematics  
University of St. Thomas  
St. Paul, MN  
USA

ISSN 0172-6056

ISSN 2197-5604 (electronic)

Undergraduate Texts in Mathematics

ISBN 978-3-319-91040-6

ISBN 978-3-319-91041-3 (eBook)

<https://doi.org/10.1007/978-3-319-91041-3>

Library of Congress Control Number: 2018941541

Mathematics Subject Classification (2010): 15-01, 15AXX, 65FXX, 05C50, 34A30, 62H25, 65D05, 65D07, 65D18

1st edition: © 2006 Pearson Education, Inc., Pearson Prentice Hall, Pearson Education, Inc., Upper Saddle River, NJ 07458

2nd edition: © Springer International Publishing AG, part of Springer Nature 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature. The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To our children and grandchildren.

You are the light of our life.

# Preface

Applied mathematics rests on two central pillars: calculus and linear algebra. While calculus has its roots in the universal laws of Newtonian physics, linear algebra arises from a much more mundane issue: the need to solve simple systems of linear algebraic equations. Despite its humble origins, linear algebra ends up playing a comparably profound role in both applied and theoretical mathematics, as well as in all of science and engineering, including computer science, data analysis and machine learning, imaging and signal processing, probability and statistics, economics, numerical analysis, mathematical biology, and many other disciplines. Nowadays, a proper grounding in both calculus and linear algebra is an essential prerequisite for a successful career in science, technology, engineering, statistics, data science, and, of course, mathematics.

Since Newton, and, to an even greater extent following Einstein, modern science has been confronted with the inherent nonlinearity of the macroscopic universe. But most of our insight and progress is based on linear approximations. Moreover, at the atomic level, quantum mechanics remains an inherently linear theory. (The complete reconciliation of linear quantum theory with the nonlinear relativistic universe remains the holy grail of modern physics.) Only with the advent of large-scale computers have we been able to begin to investigate the full complexity of natural phenomena. But computers rely on numerical algorithms, and these in turn require manipulating and solving systems of algebraic equations. Now, rather than just a handful of equations, we may be confronted by gigantic systems containing thousands (or even millions) of unknowns. Without the discipline of linear algebra to formulate systematic, efficient solution algorithms, as well as the consequent insight into how to proceed when the numerical solution is insufficiently accurate, we would be unable to make progress in the linear regime, let alone make sense of the truly nonlinear physical universe.

Linear algebra can thus be viewed as the mathematical apparatus needed to solve potentially huge linear systems, to understand their underlying structure, and to apply what is learned in other contexts. The term “linear” is the key, and, in fact, it refers not just to linear algebraic equations, but also to linear differential equations, both ordinary and partial, linear boundary value problems, linear integral equations, linear iterative systems, linear control systems, and so on. It is a profound truth that, while outwardly different, all linear systems are remarkably similar at their core. Basic mathematical principles such as linear superposition, the interplay between homogeneous and inhomogeneous systems, the Fredholm alternative characterizing solvability, orthogonality, positive definiteness and minimization principles, eigenvalues and singular values, and linear iteration, to name but a few, reoccur in surprisingly many ostensibly unrelated contexts.

In the late nineteenth and early twentieth centuries, mathematicians came to the realization that all of these disparate techniques could be subsumed in the edifice now known as linear algebra. Understanding, and, more importantly, exploiting the apparent similarities between, say, algebraic equations and differential equations, requires us to become more sophisticated — that is, more abstract — in our mode of thinking. The abstraction

process distills the essence of the problem away from all its distracting particularities, and, seen in this light, all linear systems rest on a common mathematical framework. Don't be afraid! Abstraction is not new in your mathematical education. In elementary algebra, you already learned to deal with variables, which are the abstraction of numbers. Later, the abstract concept of a function formalized particular relations between variables, say distance, velocity, and time, or mass, acceleration, and force. In linear algebra, the abstraction is raised to yet a further level, in that one views apparently different types of objects (vectors, matrices, functions, ...) and systems (algebraic, differential, integral, ...) in a common conceptual framework. (And this is by no means the end of the mathematical abstraction process; modern category theory, [37], abstractly unites different conceptual frameworks.)

In applied mathematics, we do not introduce abstraction for its intrinsic beauty. Our ultimate purpose is to develop effective methods and algorithms for applications in science, engineering, computing, statistics, data science, etc. For us, abstraction is driven by the need for understanding and insight, and is justified only if it aids in the solution to real world problems and the development of analytical and computational tools. Whereas to the beginning student the initial concepts may seem designed merely to bewilder and confuse, one must reserve judgment until genuine applications appear. Patience and perseverance are vital. Once we have acquired some familiarity with basic linear algebra, significant, interesting applications will be readily forthcoming. In this text, we encounter graph theory and networks, mechanical structures, electrical circuits, quantum mechanics, the geometry underlying computer graphics and animation, signal and image processing, interpolation and approximation, dynamical systems modeled by linear differential equations, vibrations, resonance, and damping, probability and stochastic processes, statistics, data analysis, splines and modern font design, and a range of powerful numerical solution algorithms, to name a few. Further applications of the material you learn here will appear throughout your mathematical and scientific career.

This textbook has two interrelated pedagogical goals. The first is to explain basic techniques that are used in modern, real-world problems. But we have not written a mere mathematical cookbook — a collection of linear algebraic recipes and algorithms. We believe that it is important for the applied mathematician, as well as the scientist and engineer, not just to learn mathematical techniques and how to apply them in a variety of settings, but, even more importantly, to understand why they work and how they are derived from first principles. In our approach, applications go hand in hand with theory, each reinforcing and inspiring the other. To this end, we try to lead the reader through the reasoning that leads to the important results. We do not shy away from stating theorems and writing out proofs, particularly when they lead to insight into the methods and their range of applicability. We hope to spark that eureka moment, when you realize “Yes, of course! I could have come up with that if I'd only sat down and thought it out.” Most concepts in linear algebra are not all that difficult at their core, and, by grasping their essence, not only will you know how to apply them in routine contexts, you will understand what may be required to adapt to unusual or recalcitrant problems. And, the further you go on in your studies or work, the more you realize that very few real-world problems fit neatly into the idealized framework outlined in a textbook. So it is (applied) mathematical reasoning and not mere linear algebraic technique that is the core and *raison d'être* of this text!

Applied mathematics can be broadly divided into three mutually reinforcing components. The first is modeling — how one derives the governing equations from physical

principles. The second is solution techniques and algorithms — methods for solving the model equations. The third, perhaps least appreciated but in many ways most important, are the frameworks that incorporate disparate analytical methods into a few broad themes. The key paradigms of applied linear algebra to be covered in this text include

- Gaussian Elimination and factorization of matrices;
- linearity and linear superposition;
- span, linear independence, basis, and dimension;
- inner products, norms, and inequalities;
- compatibility of linear systems via the Fredholm alternative;
- positive definiteness and minimization principles;
- orthonormality and the Gram–Schmidt process;
- least squares solutions, interpolation, and approximation;
- linear functions and linear and affine transformations;
- eigenvalues and eigenvectors/eigenfunctions;
- singular values and principal component analysis;
- linear iteration, including Markov processes and numerical solution schemes;
- linear systems of ordinary differential equations, stability, and matrix exponentials;
- vibrations, quasi-periodicity, damping, and resonance; .

These are all interconnected parts of a very general applied mathematical edifice of remarkable power and practicality. Understanding such broad themes of applied mathematics is our overarching objective. Indeed, this book began life as a part of a much larger work, whose goal is to similarly cover the full range of modern applied mathematics, both linear and nonlinear, at an advanced undergraduate level. The second installment is now in print, as the first author's text on partial differential equations, [61], which forms a natural extension of the linear analytical methods and theoretical framework developed here, now in the context of the equilibria and dynamics of continuous media, Fourier analysis, and so on. Our inspirational source was and continues to be the visionary texts of Gilbert Strang, [79, 80]. Based on students' reactions, our goal has been to present a more linearly ordered and less ambitious development of the subject, while retaining the excitement and interconnectedness of theory and applications that is evident in Strang's works.

## Syllabi and Prerequisites

This text is designed for three potential audiences:

- A beginning, in-depth course covering the fundamentals of linear algebra and its applications for highly motivated and mathematically mature students.
- A second undergraduate course in linear algebra, with an emphasis on those methods and concepts that are important in applications.
- A beginning graduate-level course in linear mathematics for students in engineering, physical science, computer science, numerical analysis, statistics, and even mathematical biology, finance, economics, social sciences, and elsewhere, as well as master's students in applied mathematics.

Although most students reading this book will have already encountered some basic linear algebra — matrices, vectors, systems of linear equations, basic solution techniques, etc. — the text makes no such assumptions. Indeed, the first chapter starts at the very beginning by introducing linear algebraic systems, matrices, and vectors, followed by very

basic Gaussian Elimination. We do assume that the reader has taken a standard two year calculus sequence. One-variable calculus — derivatives and integrals — will be used without comment; multivariable calculus will appear only fleetingly and in an inessential way. The ability to handle scalar, constant coefficient linear ordinary differential equations is also assumed, although we do briefly review elementary solution techniques in Chapter 7. Proofs by induction will be used on occasion. But the most essential prerequisite is a certain degree of mathematical maturity and willingness to handle the increased level of abstraction that lies at the heart of contemporary linear algebra.

## Survey of Topics

In addition to introducing the fundamentals of matrices, vectors, and Gaussian Elimination from the beginning, the initial chapter delves into perhaps less familiar territory, such as the (permuted)  $LU$  and  $LDV$  decompositions, and the practical numerical issues underlying the solution algorithms, thereby highlighting the computational efficiency of Gaussian Elimination coupled with Back Substitution versus methods based on the inverse matrix or determinants, as well as the use of pivoting to mitigate possibly disastrous effects of numerical round-off errors. Because the goal is to learn practical algorithms employed in contemporary applications, matrix inverses and determinants are de-emphasized — indeed, the most efficient way to compute a determinant is via Gaussian Elimination, which remains *the* key algorithm throughout the initial chapters.

Chapter 2 is the heart of linear algebra, and a successful course rests on the students' ability to assimilate the absolutely essential concepts of vector space, subspace, span, linear independence, basis, and dimension. While these ideas may well have been encountered in an introductory ordinary differential equation course, it is rare, in our experience, that students at this level are at all comfortable with them. The underlying mathematics is not particularly difficult, but enabling the student to come to grips with a new level of abstraction remains the most challenging aspect of the course. To this end, we have included a wide range of illustrative examples. Students should start by making sure they understand how a concept applies to vectors in Euclidean space  $\mathbb{R}^n$  before pressing on to less familiar territory. While one could design a course that completely avoids infinite-dimensional function spaces, we maintain that, at this level, they should be integrated into the subject right from the start. Indeed, linear analysis and applied mathematics, including Fourier methods, boundary value problems, partial differential equations, numerical solution techniques, signal processing, control theory, modern physics, especially quantum mechanics, and many, many other fields, both pure and applied, all rely on basic vector space constructions, and so learning to deal with the full range of examples is the secret to future success. Section 2.5 then introduces the fundamental subspaces associated with a matrix — kernel (null space), image (column space), coimage (row space), and cokernel (left null space) — leading to what is known as the Fundamental Theorem of Linear Algebra which highlights the remarkable interplay between a matrix and its transpose. The role of these spaces in the characterization of solutions to linear systems, e.g., the basic superposition principles, is emphasized. The final Section 2.6 covers a nice application to graph theory, in preparation for later developments.

Chapter 3 discusses general inner products and norms, using the familiar dot product and Euclidean distance as motivational examples. Again, we develop both the finite-dimensional and function space cases in tandem. The fundamental Cauchy–Schwarz inequality is easily derived in this abstract framework, and the more familiar triangle in-

equality, for norms derived from inner products, is a simple consequence. This leads to the definition of a general norm and the induced matrix norm, of fundamental importance in iteration, analysis, and numerical methods. The classification of inner products on Euclidean space leads to the important class of positive definite matrices. Gram matrices, constructed out of inner products of elements of inner product spaces, are a particularly fruitful source of positive definite and semi-definite matrices, and reappear throughout the text. Tests for positive definiteness rely on Gaussian Elimination and the connections between the  $LDL^T$  factorization of symmetric matrices and the process of completing the square in a quadratic form. We have deferred treating complex vector spaces until the final section of this chapter — only the definition of an inner product is not an evident adaptation of its real counterpart.

Chapter 4 exploits the many advantages of orthogonality. The use of orthogonal and orthonormal bases creates a dramatic speed-up in basic computational algorithms. Orthogonal matrices, constructed out of orthogonal bases, play a major role, both in geometry and graphics, where they represent rigid rotations and reflections, as well as in notable numerical algorithms. The orthogonality of the fundamental matrix subspaces leads to a linear algebraic version of the Fredholm alternative for compatibility of linear systems. We develop several versions of the basic Gram–Schmidt process for converting an arbitrary basis into an orthogonal basis, used in particular to construct orthogonal polynomials and functions. When implemented on bases of  $\mathbb{R}^n$ , the algorithm becomes the celebrated  $QR$  factorization of a nonsingular matrix. The final section surveys an important application to contemporary signal and image processing: the discrete Fourier representation of a sampled signal, culminating in the justly famous Fast Fourier Transform.

Chapter 5 is devoted to solving the most basic multivariable minimization problem: a quadratic function of several variables. The solution is reduced, by a purely algebraic computation, to a linear system, and then solved in practice by, for example, Gaussian Elimination. Applications include finding the closest element of a subspace to a given point, which is reinterpreted as the orthogonal projection of the element onto the subspace, and results in the least squares solution to an incompatible linear system. Interpolation of data points by polynomials, trigonometric function, splines, etc., and least squares approximation of discrete data and continuous functions are thereby handled in a common conceptual framework.

Chapter 6 covers some striking applications of the preceding developments in mechanics and electrical circuits. We introduce a general mathematical structure that governs a wide range of equilibrium problems. To illustrate, we start with simple mass–spring chains, followed by electrical networks, and finish by analyzing the equilibrium configurations and the stability properties of general structures. Extensions to continuous mechanical and electrical systems governed by boundary value problems for ordinary and partial differential equations can be found in the companion text [61].

Chapter 7 delves into the general abstract foundations of linear algebra, and includes significant applications to geometry. Matrices are now viewed as a particular instance of linear functions between vector spaces, which also include linear differential operators, linear integral operators, quantum mechanical operators, and so on. Basic facts about linear systems, such as linear superposition and the connections between the homogeneous and inhomogeneous systems, which were already established in the algebraic context, are shown to be of completely general applicability. Linear functions and slightly more general affine functions on Euclidean space represent basic geometrical transformations — rotations, shears, translations, screw motions, etc. — and so play an essential role in modern computer

graphics, movies, animation, gaming, design, elasticity, crystallography, symmetry, etc. Further, the elementary transpose operation on matrices is viewed as a particular case of the adjoint operation on linear functions between inner product spaces, leading to a general theory of positive definiteness that characterizes solvable quadratic minimization problems, with far-reaching consequences for modern functional analysis, partial differential equations, and the calculus of variations, all fundamental in physics and mechanics.

Chapters 8–10 are concerned with eigenvalues and their many applications, including data analysis, numerical methods, and linear dynamical systems, both continuous and discrete. After motivating the fundamental definition of eigenvalue and eigenvector through the quest to solve linear systems of ordinary differential equations, the remainder of Chapter 8 develops the basic theory and a range of applications, including eigenvector bases, diagonalization, the Schur decomposition, and the Jordan canonical form. Practical computational schemes for determining eigenvalues and eigenvectors are postponed until Chapter 9. The final two sections cover the singular value decomposition and principal component analysis, of fundamental importance in modern statistical analysis and data science.

Chapter 9 employs eigenvalues to analyze discrete dynamics, as governed by linear iterative systems. The formulation of their stability properties leads us to define the spectral radius and further develop matrix norms. Section 9.3 contains applications to Markov chains arising in probabilistic and stochastic processes. We then discuss practical alternatives to Gaussian Elimination for solving linear systems, including the iterative Jacobi, Gauss–Seidel, and Successive Over–Relaxation (SOR) schemes, as well as methods for computing eigenvalues and eigenvectors including the Power Method and its variants, and the striking  $QR$  algorithm, including a new proof of its convergence. Section 9.6 introduces more recent semi-direct iterative methods based on Krylov subspaces that are increasingly employed to solve the large sparse linear systems arising in the numerical solution of partial differential equations and elsewhere: Arnoldi and Lanczos methods, Conjugate Gradients (CG), the Full Orthogonalization Method (FOM), and the Generalized Minimal Residual Method (GMRES). The chapter concludes with a short introduction to wavelets, a powerful modern alternative to classical Fourier analysis, now used extensively throughout signal processing and imaging science.

The final Chapter 10 applies eigenvalues to linear dynamical systems modeled by systems of ordinary differential equations. After developing basic solution techniques, the focus shifts to understanding the qualitative properties of solutions and particularly the role of eigenvalues in the stability of equilibria. The two-dimensional case is discussed in full detail, culminating in a complete classification of the possible phase portraits and stability properties. Matrix exponentials are introduced as an alternative route to solving first order homogeneous systems, and are also applied to solve the inhomogeneous version, as well as to geometry, symmetry, and group theory. Our final topic is second order linear systems, which model dynamical motions and vibrations in mechanical structures and electrical circuits. In the absence of frictional damping and instabilities, solutions are quasiperiodic combinations of the normal modes. We finish by briefly discussing the effects of damping and of periodic forcing, including its potentially catastrophic role in resonance.

## Course Outlines

Our book includes far more material than can be comfortably covered in a single semester; a full year’s course would be able to do it justice. If you do not have this luxury, several

possible semester and quarter courses can be extracted from the wealth of material and applications.

First, the core of basic linear algebra that all students should know includes the following topics, which are indexed by the section numbers where they appear:

- Matrices, vectors, Gaussian Elimination, matrix factorizations, Forward and Back Substitution, inverses, determinants: 1.1–1.6, 1.8–1.9.
- Vector spaces, subspaces, linear independence, bases, dimension: 2.1–2.5.
- Inner products and their associated norms: 3.1–3.3.
- Orthogonal vectors, bases, matrices, and projections: 4.1–4.4.
- Positive definite matrices and minimization of quadratic functions: 3.4–3.5, 5.2
- Linear functions and linear and affine transformations: 7.1–7.3.
- Eigenvalues and eigenvectors: 8.2–8.3.
- Linear iterative systems: 9.1–9.2.

With these in hand, a variety of thematic threads can be extracted, including:

- Minimization, least squares, data fitting and interpolation: 4.5, 5.3–5.5.
- Dynamical systems: 8.4, 8.6 (Jordan canonical form), 10.1–10.4.
- Engineering applications: Chapter 6, 10.1–10.2, 10.5–10.6.
- Data analysis: 5.3–5.5, 8.5, 8.7–8.8.
- Numerical methods: 8.6 (Schur decomposition), 8.7, 9.1–9.2, 9.4–9.6.
- Signal processing: 3.6, 5.6, 9.7.
- Probabilistic and statistical applications: 8.7–8.8, 9.3.
- Theoretical foundations of linear algebra: Chapter 7.

For a first semester or quarter course, we recommend covering as much of the core as possible, and, if time permits, at least one of the threads, our own preference being the material on structures and circuits. One option for streamlining the syllabus is to concentrate on finite-dimensional vector spaces, bypassing the function space material, although this would deprive the students of important insight into the full scope of linear algebra.

For a second course in linear algebra, the students are typically familiar with elementary matrix methods, including the basics of matrix arithmetic, Gaussian Elimination, determinants, inverses, dot product and Euclidean norm, eigenvalues, and, often, first order systems of ordinary differential equations. Thus, much of Chapter 1 can be reviewed quickly. On the other hand, the more abstract fundamentals, including vector spaces, span, linear independence, basis, and dimension are, in our experience, still not fully mastered, and one should expect to spend a significant fraction of the early part of the course covering these essential topics from Chapter 2 in full detail. Beyond the core material, there should be time for a couple of the indicated threads depending on the audience and interest of the instructor.

Similar considerations hold for a beginning graduate level course for scientists and engineers. Here, the emphasis should be on applications required by the students, particularly numerical methods and data analysis, and function spaces should be firmly built into the class from the outset. As always, the students' mastery of the first five sections of Chapter 2 remains of paramount importance.

## Comments on Individual Chapters

*Chapter 1:* On the assumption that the students have already seen matrices, vectors, Gaussian Elimination, inverses, and determinants, most of this material will be review and should be covered at a fairly rapid pace. On the other hand, the  $LU$  decomposition and the emphasis on solution techniques centered on Forward and Back Substitution, in contrast to impractical schemes involving matrix inverses and determinants, might be new. Sections 1.7, on the practical/numerical aspects of Gaussian Elimination, is optional.

*Chapter 2:* The crux of the course. A key decision is whether to incorporate infinite-dimensional vector spaces, as is recommended and done in the text, or to have an abbreviated syllabus that covers only finite-dimensional spaces, or, even more restrictively, only  $\mathbb{R}^n$  and subspaces thereof. The last section, on graph theory, can be skipped unless you plan on covering Chapter 6 and (parts of) the final sections of Chapters 9 and 10.

*Chapter 3:* Inner products and positive definite matrices are essential, but, under time constraints, one can delay Section 3.3, on more general norms, as they begin to matter only in the later stages of Chapters 8 and 9. Section 3.6, on complex vector spaces, can be deferred until the discussions of complex eigenvalues, complex linear systems, and real and complex solutions to linear iterative and differential equations; on the other hand, it is required in Section 5.6, on discrete Fourier analysis.

*Chapter 4:* The basics of orthogonality, as covered in Sections 4.1–4.4, should be an essential part of the students' training, although one can certainly omit the final subsection in Sections 4.2 and 4.3. The final section, on orthogonal polynomials, is optional.

*Chapter 5:* We recommend covering the solution of quadratic minimization problems and at least the basics of least squares. The applications — approximation of data, interpolation and approximation by polynomials, trigonometric functions, more general functions, and splines, etc., are all optional, as is the final section on discrete Fourier methods and the Fast Fourier Transform.

*Chapter 6* provides a welcome relief from the theory for the more applied students in the class, and is one of our favorite parts to teach. While it may well be skipped, the material is particularly appealing for a class with engineering students. One could specialize to just the material on mass/spring chains and structures, or, alternatively, on electrical circuits with the connections to spectral graph theory, based on Section 2.6, and further developed in Section 8.7.

*Chapter 7:* The first third of this chapter, on linear functions, linear and affine transformations, and geometry, is part of the core. This remainder of the chapter recasts many of the linear algebraic techniques already encountered in the context of matrices and vectors in Euclidean space in a more general abstract framework, and could be skimmed over or entirely omitted if time is an issue, with the relevant constructions introduced in the context of more concrete developments, as needed.

*Chapter 8:* Eigenvalues are absolutely essential. The motivational material based on solving systems of differential equations in Section 8.1 can be skipped over. Sections 8.2 and 8.3 are the heart of the matter. Of the remaining sections, the material on symmetric matrices should have the highest priority, leading to singular values and principal component analysis and a variety of numerical methods.

*Chapter 9:* If time permits, the first two sections are well worth covering. For a numerically oriented class, Sections 9.4–9.6 would be a priority, whereas Section 9.3 studies Markov processes — an appealing probabilistic/stochastic application. The chapter concludes with an optional introduction to wavelets, which is somewhat off-topic, but nevertheless serves to combine orthogonality and iterative methods in a compelling and important modern application.

*Chapter 10* is devoted to linear systems of ordinary differential equations, their solutions, and their stability properties. The basic techniques will be a repeat to students who have already taken an introductory linear algebra and ordinary differential equations course, but the more advanced material will be new and of interest.

## Changes from the First Edition

For the Second Edition, we have revised and edited the entire manuscript, correcting all known errors and typos, and, we hope, not introducing any new ones! Some of the existing material has been rearranged. The most significant change is having moved the chapter on orthogonality to before the minimization and least squares chapter, since orthogonal vectors, bases, and subspaces, as well as the Gram–Schmidt process and orthogonal projection play an absolutely fundamental role in much of the later material. In this way, it is easier to skip over Chapter 5 with minimal loss of continuity. Matrix norms now appear much earlier in Section 3.3, since they are employed in several other locations. The second major reordering is to switch the chapters on iteration and dynamics, in that the former is more attuned to linear algebra, while the latter is oriented towards analysis. In the same vein, space constraints compelled us to delete the last chapter of the first edition, which was on boundary value problems. Although this material serves to emphasize the importance of the abstract linear algebraic techniques developed throughout the text, now extended to infinite-dimensional function spaces, the material contained therein can now all be found in the first author’s Springer Undergraduate Text in Mathematics, *Introduction to Partial Differential Equations*, [61], with the exception of the subsection on splines, which now appears at the end of Section 5.5.

There are several significant additions:

- In recognition of their increasingly essential role in modern data analysis and statistics, Section 8.7, on singular values, has been expanded, continuing into the new Section 8.8, on Principal Component Analysis, which includes a brief introduction to basic statistical data analysis.
- We have added a new Section 9.6, on Krylov subspace methods, which are increasingly employed to devise effective and efficient numerical solution schemes for sparse linear systems and eigenvalue calculations.
- Section 8.4 introduces and characterizes invariant subspaces, in recognition of their importance to dynamical systems, both finite- and infinite-dimensional, as well as linear iterative systems, and linear control systems. (Much as we would have liked also to add material on linear control theory, space constraints ultimately interfered.)
- We included some basics of spectral graph theory, of importance in contemporary theoretical computer science, data analysis, networks, imaging, etc., starting in Section 2.6 and continuing to the graph Laplacian, introduced, in the context of electrical networks, in Section 6.2, along with its spectrum — eigenvalues and singular values — in Section 8.7.

- We decided to include a short Section 9.7, on wavelets. While this perhaps fits more naturally with Section 5.6, on discrete Fourier analysis, the convergence proofs rely on the solution to an iterative linear system and hence on preceding developments in Chapter 9.
- A number of new exercises have been added, in the new sections and also scattered throughout the text.

Following the advice of friends, colleagues, and reviewers, we have also revised some of the less standard terminology used in the first edition to bring it closer to the more commonly accepted practices. Thus “range” is now “image” and “target space” is now “codomain”. The terms “special lower/upper triangular matrix” are now “lower/upper unitriangular matrix”, thus drawing attention to their unipotence. On the other hand, the term “regular” for a square matrix admitting an  $LU$  factorization has been kept, since there is really no suitable alternative appearing in the literature. Finally, we decided to retain our term “complete” for a matrix that admits a complex eigenvector basis, in lieu of “diagonalizable” (which depends upon whether one deals in the real or complex domain), “semi-simple”, or “perfect”. This choice permits us to refer to a “complete eigenvalue”, independent of the underlying status of the matrix.

## Exercises and Software

Exercises appear at the end of almost every subsection, and come in a medley of flavors. Each exercise set starts with some straightforward computational problems to test students’ comprehension and reinforce the new techniques and ideas. Ability to solve these basic problems should be thought of as a minimal requirement for learning the material. More advanced and theoretical exercises tend to appear later on in the set. Some are routine, but others are challenging computational problems, computer-based exercises and projects, details of proofs that were not given in the text, additional practical and theoretical results of interest, further developments in the subject, etc. Some will challenge even the most advanced student.

As a guide, some of the exercises are marked with special signs:

- ◊ indicates an exercise that is used at some point in the text, or is important for further development of the subject.
- ♡ indicates a project — usually an exercise with multiple interdependent parts.
- ♠ indicates an exercise that requires (or at least strongly recommends) use of a computer. The student could either be asked to write their own computer code in, say, MATLAB, MATHEMATICA, MAPLE, etc., or make use of pre-existing software packages.
- ♣ = ♠ + ♡ indicates a computer project.

*Advice to instructors:* Don’t be afraid to assign only a couple of parts of a multi-part exercise. We have found the True/False exercises to be a particularly useful indicator of a student’s level of understanding. Emphasize to the students that a full answer is not merely a T or F, but must include a detailed explanation of the reason, e.g., a proof, or a counterexample, or a reference to a result in the text, etc.

## Conventions and Notations

*Note:* A full symbol and notation index can be found at the end of the book.

Equations are numbered consecutively within chapters, so that, for example, (3.12) refers to the 12<sup>th</sup> equation in Chapter 3. Theorems, Lemmas, Propositions, Definitions, and Examples are also numbered consecutively within each chapter, using a common index. Thus, in Chapter 1, Lemma 1.2 follows Definition 1.1, and precedes Theorem 1.3 and Example 1.4. We find this numbering system to be the most conducive for navigating through the book.

References to books, papers, etc., are listed alphabetically at the end of the text, and are referred to by number. Thus, [61] indicates the 61<sup>st</sup> listed reference, which happens to be the first author's partial differential equations text.

*Q.E.D.* is placed at the end of a proof, being the abbreviation of the classical Latin phrase *quod erat demonstrandum*, which can be translated as “what was to be demonstrated”.

$\mathbb{R}, \mathbb{C}, \mathbb{Z}, \mathbb{Q}$  denote, respectively, the real numbers, the complex numbers, the integers, and the rational numbers. We use  $e \approx 2.71828182845904\dots$  to denote the base of the natural logarithm,  $\pi = 3.14159265358979\dots$  for the area of a circle of unit radius, and  $i$  to denote the imaginary unit, i.e., one of the two square roots of  $-1$ , the other being  $-i$ . The absolute value of a real number  $x$  is denoted by  $|x|$ ; more generally,  $|z|$  denotes the modulus of the complex number  $z$ .

We consistently use boldface lowercase letters, e.g.,  $\mathbf{v}, \mathbf{x}, \mathbf{a}$ , to denote vectors (almost always column vectors), whose entries are the corresponding non-bold subscripted letter:  $v_1, x_i, a_n$ , etc. Matrices are denoted by ordinary capital letters, e.g.,  $A, C, K, M$  — but not all such letters refer to matrices; for instance,  $V$  often refers to a vector space,  $L$  to a linear function, etc. The entries of a matrix, say  $A$ , are indicated by the corresponding subscripted lowercase letters,  $a_{ij}$  being the entry in its  $i^{\text{th}}$  row and  $j^{\text{th}}$  column.

We use the standard notations

$$\sum_{i=1}^n a_i = a_1 + a_2 + \cdots + a_n, \quad \prod_{i=1}^n a_i = a_1 a_2 \cdots a_n,$$

for the sum and product of the quantities  $a_1, \dots, a_n$ . We use  $\max$  and  $\min$  to denote maximum and minimum, respectively, of a closed subset of  $\mathbb{R}$ . Modular arithmetic is indicated by  $j = k \bmod n$ , for  $j, k, n \in \mathbb{Z}$  with  $n > 0$ , to mean  $j - k$  is divisible by  $n$ .

We use  $S = \{f \mid C\}$  to denote a set, where  $f$  is a formula for the members of the set and  $C$  is a list of conditions, which may be empty, in which case it is omitted. For example,  $\{x \mid 0 \leq x \leq 1\}$  means the closed unit interval from 0 to 1, also denoted  $[0, 1]$ , while  $\{ax^2 + bx + c \mid a, b, c \in \mathbb{R}\}$  is the set of real quadratic polynomials, and  $\{0\}$  is the set consisting only of the number 0. We write  $x \in S$  to indicate that  $x$  is an element of the set  $S$ , while  $y \notin S$  says that  $y$  is not an element. The cardinality, or number of elements, in the set  $A$ , which may be infinite, is denoted by  $\#A$ . The union and intersection of the sets  $A, B$  are respectively denoted by  $A \cup B$  and  $A \cap B$ . The subset notation  $A \subset B$  includes the possibility that the sets might be equal, although for emphasis we sometimes write  $A \subseteq B$ , while  $A \subsetneq B$  specifically implies that  $A \neq B$ . We can also write  $A \subset B$  as  $B \supset A$ . We use  $B \setminus A = \{x \mid x \in B, x \notin A\}$  to denote the set-theoretic difference, meaning all elements of  $B$  that do not belong to  $A$ .

An arrow  $\rightarrow$  is used in two senses: first, to indicate convergence of a sequence:  $x_n \rightarrow x^*$  as  $n \rightarrow \infty$ ; second, to indicate a function, so  $f: X \rightarrow Y$  means that  $f$  defines a function from the domain set  $X$  to the codomain set  $Y$ , written  $y = f(x) \in Y$  for  $x \in X$ . We use  $\equiv$  to emphasize when two functions agree everywhere, so  $f(x) \equiv 1$  means that  $f$  is the constant function, equal to 1 at all values of  $x$ . Composition of functions is denoted  $f \circ g$ .

Angles are always measured in radians (although occasionally degrees will be mentioned in descriptive sentences). All trigonometric functions,  $\cos$ ,  $\sin$ ,  $\tan$ ,  $\sec$ , etc., are evaluated on radians. (Make sure your calculator is locked in radian mode!)

As usual, we denote the natural exponential function by  $e^x$ . We always use  $\log x$  for its inverse — the natural (base  $e$ ) logarithm (never the ugly modern version  $\ln x$ ), while  $\log_a x = \log x / \log a$  is used for logarithms with base  $a$ .

We follow the reference tome [59] (whose mathematical editor is the first author's father) and use  $\text{ph } z$  for the phase of a complex number. We prefer this to the more common term “argument”, which is also used to refer to the argument of a function  $f(z)$ , while “phase” is completely unambiguous and hence to be preferred.

We will employ a variety of standard notations for derivatives. In the case of ordinary derivatives, the most basic is the Leibnizian notation  $\frac{du}{dx}$  for the derivative of  $u$  with respect to  $x$ ; an alternative is the Lagrangian prime notation  $u'$ . Higher order derivatives are similar, with  $u''$  denoting  $\frac{d^2u}{dx^2}$ , while  $u^{(n)}$  denotes the  $n^{\text{th}}$  order derivative  $\frac{d^n u}{dx^n}$ . If the function depends on time,  $t$ , instead of space,  $x$ , then we use the Newtonian dot notation,  $\dot{u} = \frac{du}{dt}$ ,  $\ddot{u} = \frac{d^2u}{dt^2}$ . We use the full Leibniz notation  $\frac{\partial u}{\partial x}$ ,  $\frac{\partial u}{\partial t}$ ,  $\frac{\partial^2 u}{\partial x^2}$ ,  $\frac{\partial^2 u}{\partial x \partial t}$ , for partial derivatives of functions of several variables. All functions are assumed to be sufficiently smooth that any indicated derivatives exist and mixed partial derivatives are equal, cf. [2].

Definite integrals are denoted by  $\int_a^b f(x) dx$ , while  $\int f(x) dx$  is the corresponding indefinite integral or anti-derivative. In general, limits are denoted by  $\lim_{x \rightarrow y}$ , while  $\lim_{x \rightarrow y^+}$  and  $\lim_{x \rightarrow y^-}$  are used to denote the two one-sided limits in  $\mathbb{R}$ .

## History and Biography

Mathematics is both a historical and a social activity, and many of the algorithms, theorems, and formulas are named after famous (and, on occasion, not-so-famous) mathematicians, scientists, engineers, etc. — usually, but not necessarily, the one(s) who first came up with the idea. We try to indicate first names, approximate dates, and geographic locations of most of the named contributors. Readers who are interested in additional historical details, complete biographies, and, when available, portraits or photos, are urged to consult the wonderful University of St. Andrews MacTutor History of Mathematics archive:

<http://www-history.mcs.st-and.ac.uk>

## Some Final Remarks

*To the student:* You are about to learn modern applied linear algebra. We hope you enjoy the experience and profit from it in your future studies and career. (Indeed, we recommended holding onto this book to use for future reference.) Please send us your comments, suggestions for improvement, along with any errors you might spot. Did you find our explanations helpful or confusing? Were enough examples included in the text? Were the exercises of sufficient variety and at an appropriate level to enable you to learn the material?

*To the instructor:* Thank you for adopting our text! We hope you enjoy teaching from it as much as we enjoyed writing it. Whatever your experience, we want to hear from you. Let us know which parts you liked and which you didn't. Which sections worked and which were less successful. Which parts your students enjoyed, which parts they struggled with, and which parts they disliked. How can we improve it?

Like every author, we sincerely hope that we have written an error-free text. Indeed, all known errors in the first edition have been corrected here. On the other hand, judging from experience, we know that, no matter how many times you proofread, mistakes still manage to sneak through. So we ask your indulgence to correct the few (we hope) that remain. Even better, email us with your questions, typos, mathematical errors and obscurities, comments, suggestions, etc.

The second edition's dedicated web site

<http://www.math.umn.edu/~olver/ala2.html>

will contain a list of known errors, commentary, feedback, and resources, as well as a number of illustrative MATLAB programs that we've used when teaching the course. Links to the *Selected Solutions Manual* will also be posted there.

## Acknowledgments

First, let us express our profound gratitude to Gil Strang for his continued encouragement from the very beginning of this undertaking. Readers familiar with his groundbreaking texts and remarkable insight can readily find his influence throughout our book. We thank Pavel Belik, Tim Garoni, Donald Kahn, Markus Keel, Cristina Santa Marta, Nilima Nigam, Greg Pierce, Fadil Santosa, Wayne Schmaedeke, Jackie Shen, Peter Shook, Thomas Scofield, and Richard Varga, as well as our classes and students, particularly Tatjala Carvalho, Colleen Duffy, and Ryan Lloyd, and last, but certainly not least, our late father/father-in-law Frank W.J. Olver and son Sheehan Olver, for proofreading, corrections, remarks, and useful suggestions that helped us create the first edition. We acknowledge Mikhail Shvartsman's contributions to the arduous task of writing out the solutions manual. We also acknowledge the helpful feedback from the reviewers of the original manuscript: Augustin Banyaga, Robert Cramer, James Curry, Jerome Dancis, Bruno Harris, Norman Johnson, Cerry Klein, Doron Lubinsky, Juan Manfredi, Fabio Augusto Milner, Tzuong-Tsieng Moh, Paul S. Muhly, Juan Carlos Álvarez Paiva, John F. Rossi, Brian Shader, Shagi-Di Shih, Tamas Wiandt, and two anonymous reviewers.

We thank many readers and students for their strongly encouraging remarks, that cumulatively helped inspire us to contemplate making this new edition. We would particularly like to thank Nihat Bayhan, Joe Benson, James Broomfield, Juan Cockburn, Richard Cook, Stephen DeSalvo, Anne Dougherty, Ken Driessel, Kathleen Fuller, Mary Halloran, Stuart Hastings, David Hiebeler, Jeffrey Humpherys, Roberta Jaskolski, Tian-Jun Li, James Meiss, Willard Miller, Jr., Sean Rostami, Arnd Scheel, Timo Schürg, David Tieri, Peter Webb, Timothy Welle, and an anonymous reviewer for their comments on, suggestions for, and corrections to the three printings of the first edition that have led to this improved second edition. We particularly want to thank Linda Ness for extensive help with the sections on SVD and PCA, including suggestions for some of the exercises. We also thank David Kramer for his meticulous proofreading of the text.

And of course, we owe an immense debt to Loretta Bartolini and Achi Dosanjh at Springer, first for encouraging us to take on a second edition, and then for their willingness to work with us to produce the book you now have in hand — especially Loretta's unwavering support, patience, and advice during the preparation of the manuscript, including encouraging us to adopt and helping perfect the full-color layout, which we hope you enjoy.

Peter J. Olver  
University of Minnesota  
[olver@umn.edu](mailto:olver@umn.edu)

Cheri Shakiban  
University of St. Thomas  
[cshakiban@stthomas.edu](mailto:cshakiban@stthomas.edu)

Minnesota, March 2018

# Table of Contents

Preface . . . . .	vii
Chapter 1. Linear Algebraic Systems . . . . .	1
1.1. Solution of Linear Systems . . . . .	1
1.2. Matrices and Vectors . . . . .	3
Matrix Arithmetic . . . . .	5
1.3. Gaussian Elimination — Regular Case . . . . .	12
Elementary Matrices . . . . .	16
The <i>LU</i> Factorization . . . . .	18
Forward and Back Substitution . . . . .	20
1.4. Pivoting and Permutations . . . . .	22
Permutations and Permutation Matrices . . . . .	25
The Permuted <i>LU</i> Factorization . . . . .	27
1.5. Matrix Inverses . . . . .	31
Gauss–Jordan Elimination . . . . .	35
Solving Linear Systems with the Inverse . . . . .	40
The <i>LDV</i> Factorization . . . . .	41
1.6. Transposes and Symmetric Matrices . . . . .	43
Factorization of Symmetric Matrices . . . . .	45
1.7. Practical Linear Algebra . . . . .	48
Tridiagonal Matrices . . . . .	52
Pivoting Strategies . . . . .	55
1.8. General Linear Systems . . . . .	59
Homogeneous Systems . . . . .	67
1.9. Determinants . . . . .	69
Chapter 2. Vector Spaces and Bases . . . . .	75
2.1. Real Vector Spaces . . . . .	76
2.2. Subspaces . . . . .	81
2.3. Span and Linear Independence . . . . .	87
Linear Independence and Dependence . . . . .	92
2.4. Basis and Dimension . . . . .	98
2.5. The Fundamental Matrix Subspaces . . . . .	105
Kernel and Image . . . . .	105
The Superposition Principle . . . . .	110
Adjoint Systems, Cokernel, and Coimage . . . . .	112
The Fundamental Theorem of Linear Algebra . . . . .	114
2.6. Graphs and Digraphs . . . . .	120

Chapter 3. Inner Products and Norms . . . . .	129
3.1. Inner Products . . . . .	129
Inner Products on Function Spaces . . . . .	133
3.2. Inequalities . . . . .	137
The Cauchy–Schwarz Inequality . . . . .	137
Orthogonal Vectors . . . . .	140
The Triangle Inequality . . . . .	142
3.3. Norms . . . . .	144
Unit Vectors . . . . .	148
Equivalence of Norms . . . . .	150
Matrix Norms . . . . .	153
3.4. Positive Definite Matrices . . . . .	156
Gram Matrices . . . . .	161
3.5. Completing the Square . . . . .	166
The Cholesky Factorization . . . . .	171
3.6. Complex Vector Spaces . . . . .	172
Complex Numbers . . . . .	173
Complex Vector Spaces and Inner Products . . . . .	177
Chapter 4. Orthogonality . . . . .	183
4.1. Orthogonal and Orthonormal Bases . . . . .	184
Computations in Orthogonal Bases . . . . .	188
4.2. The Gram–Schmidt Process . . . . .	192
Modifications of the Gram–Schmidt Process . . . . .	197
4.3. Orthogonal Matrices . . . . .	200
The $QR$ Factorization . . . . .	205
Ill-Conditioned Systems and Householder’s Method . . . . .	208
4.4. Orthogonal Projections and Orthogonal Subspaces . . . . .	212
Orthogonal Projection . . . . .	213
Orthogonal Subspaces . . . . .	216
Orthogonality of the Fundamental Matrix Subspaces and the Fredholm Alternative . . . . .	221
4.5. Orthogonal Polynomials . . . . .	226
The Legendre Polynomials . . . . .	227
Other Systems of Orthogonal Polynomials . . . . .	231
Chapter 5. Minimization and Least Squares . . . . .	235
5.1. Minimization Problems . . . . .	235
Equilibrium Mechanics . . . . .	236
Solution of Equations . . . . .	236
The Closest Point . . . . .	238
5.2. Minimization of Quadratic Functions . . . . .	239
5.3. The Closest Point . . . . .	245
5.4. Least Squares . . . . .	250

5.5. Data Fitting and Interpolation . . . . .	254
Polynomial Approximation and Interpolation . . . . .	259
Approximation and Interpolation by General Functions . . . . .	271
Least Squares Approximation in Function Spaces . . . . .	274
Orthogonal Polynomials and Least Squares . . . . .	277
Splines . . . . .	279
5.6. Discrete Fourier Analysis and the Fast Fourier Transform . . . . .	285
Compression and Denoising . . . . .	293
The Fast Fourier Transform . . . . .	295
 Chapter 6. Equilibrium . . . . .	301
6.1. Springs and Masses . . . . .	301
Positive Definiteness and the Minimization Principle . . . . .	309
6.2. Electrical Networks . . . . .	311
Batteries, Power, and the Electrical–Mechanical Correspondence .	317
6.3. Structures . . . . .	322
 Chapter 7. Linearity . . . . .	341
7.1. Linear Functions . . . . .	342
Linear Operators . . . . .	347
The Space of Linear Functions . . . . .	349
Dual Spaces . . . . .	350
Composition . . . . .	352
Inverses . . . . .	355
7.2. Linear Transformations . . . . .	358
Change of Basis . . . . .	365
7.3. Affine Transformations and Isometries . . . . .	370
Isometry . . . . .	372
7.4. Linear Systems . . . . .	376
The Superposition Principle . . . . .	378
Inhomogeneous Systems . . . . .	383
Superposition Principles for Inhomogeneous Systems . . . . .	388
Complex Solutions to Real Systems . . . . .	390
7.5. Adjoint, Positive Definite Operators, and Minimization Principles .	395
Self-Adjoint and Positive Definite Linear Functions . . . . .	398
Minimization . . . . .	400
 Chapter 8. Eigenvalues and Singular Values . . . . .	403
8.1. Linear Dynamical Systems . . . . .	404
Scalar Ordinary Differential Equations . . . . .	404
First Order Dynamical Systems . . . . .	407
8.2. Eigenvalues and Eigenvectors . . . . .	408
Basic Properties of Eigenvalues . . . . .	415
The Gershgorin Circle Theorem . . . . .	420

8.3. Eigenvector Bases . . . . .	423
Diagonalization . . . . .	426
8.4. Invariant Subspaces . . . . .	429
8.5. Eigenvalues of Symmetric Matrices . . . . .	431
The Spectral Theorem . . . . .	437
Optimization Principles for Eigenvalues of Symmetric Matrices . .	440
8.6. Incomplete Matrices . . . . .	444
The Schur Decomposition . . . . .	444
The Jordan Canonical Form . . . . .	447
8.7. Singular Values . . . . .	454
The Pseudoinverse . . . . .	457
The Euclidean Matrix Norm . . . . .	459
Condition Number and Rank . . . . .	460
Spectral Graph Theory . . . . .	462
8.8. Principal Component Analysis . . . . .	467
Variance and Covariance . . . . .	467
The Principal Components . . . . .	471
 Chapter 9. Iteration . . . . .	475
9.1. Linear Iterative Systems . . . . .	476
Scalar Systems . . . . .	476
Powers of Matrices . . . . .	479
Diagonalization and Iteration . . . . .	484
9.2. Stability . . . . .	488
Spectral Radius . . . . .	489
Fixed Points . . . . .	493
Matrix Norms and Convergence . . . . .	495
9.3. Markov Processes . . . . .	499
9.4. Iterative Solution of Linear Algebraic Systems . . . . .	506
The Jacobi Method . . . . .	508
The Gauss–Seidel Method . . . . .	512
Successive Over-Relaxation . . . . .	517
9.5. Numerical Computation of Eigenvalues . . . . .	522
The Power Method . . . . .	522
The $QR$ Algorithm . . . . .	526
Tridiagonalization . . . . .	532
9.6. Krylov Subspace Methods . . . . .	536
Krylov Subspaces . . . . .	536
Arnoldi Iteration . . . . .	537
The Full Orthogonalization Method . . . . .	540
The Conjugate Gradient Method . . . . .	542
The Generalized Minimal Residual Method . . . . .	546

9.7. Wavelets . . . . .	549
The Haar Wavelets . . . . .	549
Modern Wavelets . . . . .	555
Solving the Dilation Equation . . . . .	559
 Chapter 10. Dynamics . . . . .	565
10.1. Basic Solution Techniques . . . . .	565
The Phase Plane . . . . .	567
Existence and Uniqueness . . . . .	570
Complete Systems . . . . .	572
The General Case . . . . .	575
10.2. Stability of Linear Systems . . . . .	579
10.3. Two-Dimensional Systems . . . . .	585
Distinct Real Eigenvalues . . . . .	586
Complex Conjugate Eigenvalues . . . . .	587
Incomplete Double Real Eigenvalue . . . . .	588
Complete Double Real Eigenvalue . . . . .	588
10.4. Matrix Exponentials . . . . .	592
Applications in Geometry . . . . .	599
Invariant Subspaces and Linear Dynamical Systems . . . . .	603
Inhomogeneous Linear Systems . . . . .	605
10.5. Dynamics of Structures . . . . .	608
Stable Structures . . . . .	610
Unstable Structures . . . . .	615
Systems with Differing Masses . . . . .	618
Friction and Damping . . . . .	620
10.6. Forcing and Resonance . . . . .	623
Electrical Circuits . . . . .	628
Forcing and Resonance in Systems . . . . .	630
 References . . . . .	633
Symbol Index . . . . .	637
Subject Index . . . . .	643



# Chapter 1

# Linear Algebraic Systems

Linear algebra is the core of modern applied mathematics. Its humble origins are to be found in the need to solve “elementary” systems of linear algebraic equations. But its ultimate scope is vast, impinging on all of mathematics, both pure and applied, as well as numerical analysis, statistics, data science, physics, engineering, mathematical biology, financial mathematics, and every other discipline in which mathematical methods are required. A thorough grounding in the methods and theory of linear algebra is an essential prerequisite for understanding and harnessing the power of mathematics throughout its multifaceted applications.

In the first chapter, our focus will be on the most basic method for solving linear algebraic systems, known as *Gaussian Elimination* in honor of one of the all-time mathematical greats, the early nineteenth-century German mathematician Carl Friedrich Gauss, although the method appears in Chinese mathematical texts from around 150 CE, if not earlier, and was also known to Isaac Newton. Gaussian Elimination is quite elementary, but remains one of the most important algorithms in applied (as well as theoretical) mathematics. Our initial focus will be on the most important class of systems: those involving the same number of equations as unknowns — although we will eventually develop techniques for handling completely general linear systems. While the former typically have a unique solution, general linear systems may have either no solutions or infinitely many solutions. Since physical models require existence and uniqueness of their solution, the systems arising in applications often (but not always) involve the same number of equations as unknowns. Nevertheless, the ability to confidently handle all types of linear systems is a basic prerequisite for further progress in the subject. In contemporary applications, particularly those arising in numerical solutions of differential equations, in signal and image processing, and in contemporary data analysis, the governing linear systems can be huge, sometimes involving millions of equations in millions of unknowns, challenging even the most powerful supercomputer. So, a systematic and careful development of solution techniques is essential. Section 1.7 discusses some of the practical issues and limitations in computer implementations of the Gaussian Elimination method for large systems arising in applications.

Modern linear algebra relies on the basic concepts of scalar, vector, and matrix, and so we must quickly review the fundamentals of matrix arithmetic. Gaussian Elimination can be profitably reinterpreted as a certain matrix factorization, known as the (permuted) *LU* decomposition, which provides valuable insight into the solution algorithms. Matrix inverses and determinants are also discussed in brief, primarily for their theoretical properties. As we shall see, formulas relying on the inverse or the determinant are extremely inefficient, and so, except in low-dimensional or highly structured environments, are to be avoided in almost all practical computations. In the theater of applied linear algebra, Gaussian Elimination and matrix factorization are the stars, while inverses and determinants are relegated to the supporting cast.

## 1.1 Solution of Linear Systems

Gaussian Elimination is a simple, systematic algorithm to solve systems of linear equations. It is the workhorse of linear algebra, and, as such, of absolutely fundamental importance

in applied mathematics. In this section, we review the method in the most important case, in which there is the same number of equations as unknowns. The general situation will be deferred until Section 1.8.

To illustrate, consider an elementary system of three linear equations

$$\begin{aligned} x + 2y + z &= 2, \\ 2x + 6y + z &= 7, \\ x + y + 4z &= 3, \end{aligned} \tag{1.1}$$

in three unknowns  $x, y, z$ . Linearity<sup>†</sup> refers to the fact that the unknowns only appear to the first power, and there are no product terms like  $xy$  or  $xyz$ . The basic solution method is to systematically employ the following fundamental operation:

**Linear System Operation #1:** Add a multiple of one equation to another equation.

Before continuing, you might try to convince yourself that this operation doesn't change the solutions to the system. Our goal is to judiciously apply the operation and so be led to a much simpler linear system that is easy to solve, and, moreover, has the same solutions as the original. Any linear system that is derived from the original system by successive application of such operations will be called an *equivalent system*. By the preceding remark, *equivalent linear systems have the same solutions*.

The systematic feature is that we successively eliminate the variables in our equations in order of appearance. We begin by eliminating the first variable,  $x$ , from the second equation. To this end, we subtract twice the first equation from the second, leading to the equivalent system

$$\begin{aligned} x + 2y + z &= 2, \\ 2y - z &= 3, \\ x + y + 4z &= 3. \end{aligned} \tag{1.2}$$

Next, we eliminate  $x$  from the third equation by subtracting the first equation from it:

$$\begin{aligned} x + 2y + z &= 2, \\ 2y - z &= 3, \\ -y + 3z &= 1. \end{aligned} \tag{1.3}$$

The equivalent system (1.3) is already simpler than the original (1.1). Notice that the second and third equations do not involve  $x$  (by design) and so constitute a system of two linear equations for two unknowns. Moreover, once we have solved this subsystem for  $y$  and  $z$ , we can substitute the answer into the first equation, and we need only solve a single linear equation for  $x$ .

We continue on in this fashion, the next phase being the elimination of the second variable,  $y$ , from the third equation by adding  $\frac{1}{2}$  the second equation to it. The result is

$$\begin{aligned} x + 2y + z &= 2, \\ 2y - z &= 3, \\ \frac{5}{2}z &= \frac{5}{2}, \end{aligned} \tag{1.4}$$

which is the simple system we are after. It is in what is called *triangular form*, which means that, while the first equation involves all three variables, the second equation involves only the second and third variables, and the last equation involves only the last variable.

---

<sup>†</sup> The “official” definition of linearity will be deferred until Chapter 7.

Any triangular system can be straightforwardly solved by the method of *Back Substitution*. As the name suggests, we work backwards, solving the last equation first, which requires that  $z = 1$ . We substitute this result back into the penultimate equation, which becomes  $2y - 1 = 3$ , with solution  $y = 2$ . We finally substitute these two values for  $y$  and  $z$  into the first equation, which becomes  $x + 5 = 2$ , and so the solution to the triangular system (1.4) is

$$x = -3, \quad y = 2, \quad z = 1. \quad (1.5)$$

Moreover, since we used only our basic linear system operation to pass from (1.1) to the triangular system (1.4), this is also the solution to the original system of linear equations, as you can check. We note that the system (1.1) has a unique — meaning one and only one — solution, namely (1.5).

And that, barring a few minor complications that can crop up from time to time, is all that there is to the method of Gaussian Elimination! It is extraordinarily simple, but its importance cannot be overemphasized. Before exploring the relevant issues, it will help to reformulate our method in a more convenient matrix notation.

## Exercises

1.1.1. Solve the following systems of linear equations by reducing to triangular form and then using Back Substitution.

$$\begin{array}{llll} (a) \begin{array}{l} x - y = 7, \\ x + 2y = 3; \end{array} & (b) \begin{array}{l} 6u + v = 5, \\ 3u - 2v = 5; \end{array} & (c) \begin{array}{l} 2p - q + 3r = 3, \\ -p - q = 6; \end{array} & (d) \begin{array}{l} 2u - v + 2w = 2, \\ -u - v + 3w = 1, \\ 3u - 2w = 1; \end{array} \\ (e) \begin{array}{l} 5x_1 + 3x_2 - x_3 = 9, \\ 3x_1 + 2x_2 - x_3 = 5, \\ x_1 + x_2 + x_3 = -1; \end{array} & (f) \begin{array}{l} x + z - 2w = -3, \\ 2x - y + 2z - w = -5, \\ -6y - 4z + 2w = 2, \\ x + 3y + 2z - w = 1; \end{array} & (g) \begin{array}{l} 3x_1 + x_2 = 1, \\ x_1 + 3x_2 + x_3 = 1, \\ x_2 + 3x_3 + x_4 = 1, \\ x_3 + 3x_4 = 1. \end{array} \end{array}$$

1.1.2. How should the coefficients  $a$ ,  $b$ , and  $c$  be chosen so that the system  $ax + by + cz = 3$ ,  $ax - y + cz = 1$ ,  $x + by - cz = 2$ , has the solution  $x = 1$ ,  $y = 2$  and  $z = -1$ ?

♡ 1.1.3. The system  $2x = -6$ ,  $-4x + 3y = 3$ ,  $x + 4y - z = 7$ , is in *lower triangular form*.

- (a) Formulate a method of *Forward Substitution* to solve it. (b) What happens if you reduce the system to (upper) triangular form using the algorithm in this section? (c) Devise an algorithm that uses our linear system operation to reduce a system to lower triangular form and then solve it by Forward Substitution. (d) Check your algorithm by applying it to one or two of the systems in Exercise 1.1.1. Are you able to solve them in all cases?

## 1.2 Matrices and Vectors

A *matrix* is a rectangular array of numbers. Thus,

$$\begin{pmatrix} 1 & 0 & 3 \\ -2 & 4 & 1 \end{pmatrix}, \quad \begin{pmatrix} \pi & 0 \\ e & \frac{1}{2} \\ -1 & .83 \\ \sqrt{5} & -\frac{4}{7} \end{pmatrix}, \quad (.2 \quad -1.6 \quad .32), \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 3 \\ -2 & 5 \end{pmatrix},$$

are all examples of matrices. We use the notation

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad (1.6)$$

for a general matrix of size  $m \times n$  (read “ $m$  by  $n$ ”), where  $m$  denotes the number of *rows* in  $A$  and  $n$  denotes the number of *columns*. Thus, the preceding examples of matrices have respective sizes  $2 \times 3$ ,  $4 \times 2$ ,  $1 \times 3$ ,  $2 \times 1$ , and  $2 \times 2$ . A matrix is *square* if  $m = n$ , i.e., it has the same number of rows as columns. A *column vector* is an  $m \times 1$  matrix, while a *row vector* is a  $1 \times n$  matrix. As we shall see, column vectors are by far the more important of the two, and the term “vector” without qualification will always mean “column vector”. A  $1 \times 1$  matrix, which has but a single entry, is both a row and a column vector.

The number that lies in the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column of  $A$  is called the  $(i, j)$  *entry* of  $A$ , and is denoted by  $a_{ij}$ . The row index always appears first and the column index second.<sup>†</sup> Two matrices are equal,  $A = B$ , if and only if they have the same size, say  $m \times n$ , and *all* their entries are the same:  $a_{ij} = b_{ij}$  for  $i = 1, \dots, m$  and  $j = 1, \dots, n$ .

A general linear system of  $m$  equations in  $n$  unknowns will take the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\vdots && \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \quad (1.7)$$

As such, it is composed of three basic ingredients: the  $m \times n$  *coefficient matrix*  $A$ , with

entries  $a_{ij}$  as in (1.6), the column vector  $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$  containing the *unknowns*, and the column vector  $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$  containing *right-hand sides*. In our previous example,

$$\begin{aligned} x + 2y + z &= 2, \\ 2x + 6y + z &= 7, \quad \text{the coefficient matrix } A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} \text{ can be filled in, entry by entry,} \\ x + y + 4z &= 3, \end{aligned}$$

from the coefficients of the variables appearing in the equations; if a variable does not

appear in an equation, the corresponding matrix entry is 0. The vector  $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$  lists the variables, while the entries of  $\mathbf{b} = \begin{pmatrix} 2 \\ 7 \\ 3 \end{pmatrix}$  are the right-hand sides of the equations.

---

<sup>†</sup> In tensor analysis, [1], a sub- and super-script notation is adopted, with  $a_j^i$  denoting the  $(i, j)$  entry of the matrix  $A$ . This has certain advantages, but, to avoid possible confusion with powers, we shall stick with the simpler subscript notation throughout this text.

**Remark.** We will consistently use bold face lower case letters to denote vectors, and ordinary capital letters to denote general matrices.

---

## Exercises

- 1.2.1. Let  $A = \begin{pmatrix} -2 & 0 & 1 & 3 \\ -1 & 2 & 7 & -5 \\ 6 & -6 & -3 & 4 \end{pmatrix}$ . (a) What is the size of  $A$ ? (b) What is its  $(2, 3)$  entry? (c)  $(3, 1)$  entry? (d) 1<sup>st</sup> row? (e) 2<sup>nd</sup> column?
- 1.2.2. Write down examples of (a) a  $3 \times 3$  matrix; (b) a  $2 \times 3$  matrix; (c) a matrix with 3 rows and 4 columns; (d) a row vector with 4 entries; (e) a column vector with 3 entries; (f) a matrix that is both a row vector and a column vector.
- 1.2.3. For which values of  $x, y, z, w$  are the matrices  $\begin{pmatrix} x+y & x-z \\ y+w & x+2w \end{pmatrix}$  and  $\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$  equal?
- 1.2.4. For each of the systems in Exercise 1.1.1, write down the coefficient matrix  $A$  and the vectors  $\mathbf{x}$  and  $\mathbf{b}$ .
- 1.2.5. Write out and solve the linear systems corresponding to the indicated matrix, vector of unknowns, and right-hand side. (a)  $A = \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}$ ,  $\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} -1 \\ -3 \end{pmatrix}$ ; (b)  $A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ ,  $\mathbf{x} = \begin{pmatrix} u \\ v \\ w \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}$ ; (c)  $A = \begin{pmatrix} 3 & 0 & -1 \\ -2 & -1 & 0 \\ 1 & 1 & -3 \end{pmatrix}$ ,  $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ ; (d)  $A = \begin{pmatrix} 1 & 1 & -1 & -1 \\ -1 & 0 & 1 & 2 \\ 1 & -1 & 1 & 0 \\ 0 & 2 & -1 & 1 \end{pmatrix}$ ,  $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 0 \\ 4 \\ 1 \\ 5 \end{pmatrix}$ .

---

## Matrix Arithmetic

Matrix arithmetic involves three basic operations: *matrix addition*, *scalar multiplication*, and *matrix multiplication*. First we define *addition* of matrices. You are allowed to add two matrices only if they are of the *same size*, and matrix addition is performed entry by entry. For example,

$$\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} + \begin{pmatrix} 3 & -5 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 4 & -3 \\ 1 & 1 \end{pmatrix}.$$

Therefore, if  $A$  and  $B$  are  $m \times n$  matrices, their sum  $C = A + B$  is the  $m \times n$  matrix whose entries are given by  $c_{ij} = a_{ij} + b_{ij}$  for  $i = 1, \dots, m$  and  $j = 1, \dots, n$ . When defined, matrix addition is commutative,  $A + B = B + A$ , and associative,  $A + (B + C) = (A + B) + C$ , just like ordinary addition.

A *scalar* is a fancy name for an ordinary number — the term merely distinguishes it from a vector or a matrix. For the time being, we will restrict our attention to real scalars and matrices with real entries, but eventually complex scalars and complex matrices must be dealt with. We will consistently identify a scalar  $c \in \mathbb{R}$  with the  $1 \times 1$  matrix  $(c)$  in which it is the sole entry, and so will omit the redundant parentheses in the latter case. *Scalar multiplication* takes a scalar  $c$  and an  $m \times n$  matrix  $A$  and computes the  $m \times n$

matrix  $B = cA$  by multiplying each entry of  $A$  by  $c$ . For example,

$$3 \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ -3 & 0 \end{pmatrix}.$$

In general,  $b_{ij} = ca_{ij}$  for  $i = 1, \dots, m$  and  $j = 1, \dots, n$ . Basic properties of scalar multiplication are summarized at the end of this section.

Finally, we define *matrix multiplication*. First, the product of a row vector  $\mathbf{a}$  and a column vector  $\mathbf{x}$  having the *same* number of entries is the *scalar* or  $1 \times 1$  matrix defined by the following rule:

$$\mathbf{a} \mathbf{x} = (a_1 \ a_2 \ \dots \ a_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1 x_1 + a_2 x_2 + \dots + a_n x_n = \sum_{k=1}^n a_k x_k. \quad (1.8)$$

More generally, if  $A$  is an  $m \times n$  matrix and  $B$  is an  $n \times p$  matrix, so that the number of *columns* in  $A$  equals the number of *rows* in  $B$ , then the matrix product  $C = AB$  is defined as the  $m \times p$  matrix whose  $(i, j)$  entry equals the vector product of the  $i^{\text{th}}$  row of  $A$  and the  $j^{\text{th}}$  column of  $B$ . Therefore,

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}. \quad (1.9)$$

Note that our restriction on the sizes of  $A$  and  $B$  guarantees that the relevant row and column vectors will have the same number of entries, and so their product is defined.

For example, the product of the coefficient matrix  $A$  and vector of unknowns  $\mathbf{x}$  for our original system (1.1) is given by

$$A \mathbf{x} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y + z \\ 2x + 6y + z \\ x + y + 4z \end{pmatrix}.$$

The result is a column vector whose entries reproduce the left-hand sides of the original linear system! As a result, we can rewrite the system

$$A \mathbf{x} = \mathbf{b} \quad (1.10)$$

as an equality between two column vectors. This result is general; a linear system (1.7) consisting of  $m$  equations in  $n$  unknowns can be written in the matrix form (1.10), where  $A$  is the  $m \times n$  coefficient matrix (1.6),  $\mathbf{x}$  is the  $n \times 1$  column vector of unknowns, and  $\mathbf{b}$  is the  $m \times 1$  column vector containing the right-hand sides. This is one of the principal reasons for the non-evident definition of matrix multiplication. Component-wise multiplication of matrix entries turns out to be almost completely useless in applications.

Now, the bad news. Matrix multiplication is *not* commutative — that is,  $BA$  is not necessarily equal to  $AB$ . For example,  $BA$  may not be defined even when  $AB$  is. Even if both are defined, they may be different sized matrices. For example the product  $s = \mathbf{r} \mathbf{c}$  of a row vector  $\mathbf{r}$ , a  $1 \times n$  matrix, and a column vector  $\mathbf{c}$ , an  $n \times 1$  matrix with the same number of entries, is a  $1 \times 1$  matrix, or scalar, whereas the reversed product  $C = \mathbf{c} \mathbf{r}$  is an  $n \times n$  matrix. For instance,

$$(1 \ 2) \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 3, \quad \text{whereas} \quad \begin{pmatrix} 3 \\ 0 \end{pmatrix} (1 \ 2) = \begin{pmatrix} 3 & 6 \\ 0 & 0 \end{pmatrix}.$$

In computing the latter product, don't forget that we multiply the *rows* of the first matrix by the *columns* of the second, each of which has but a single entry. Moreover, even if the matrix products  $AB$  and  $BA$  have the same size, which requires both  $A$  and  $B$  to be square matrices, we may still have  $AB \neq BA$ . For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} -2 & 5 \\ -4 & 11 \end{pmatrix} \neq \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

On the other hand, matrix multiplication is associative, so  $A(BC) = (AB)C$  whenever  $A$  has size  $m \times n$ ,  $B$  has size  $n \times p$ , and  $C$  has size  $p \times q$ ; the result is a matrix of size  $m \times q$ . The proof of associativity is a tedious computation based on the definition of matrix multiplication that, for brevity, we omit.<sup>†</sup> Consequently, the one difference between matrix algebra and ordinary algebra is that you need to be careful not to change the order of multiplicative factors without proper justification.

Since matrix multiplication acts by multiplying rows by columns, one can compute the columns in a matrix product  $AB$  by multiplying the matrix  $A$  and the individual columns of  $B$ . For example, the two columns of the matrix product

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ 0 & 2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ 8 & 6 \end{pmatrix}$$

are obtained by multiplying the first matrix with the individual columns of the second:

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

In general, if we use  $\mathbf{b}_k$  to denote the  $k^{\text{th}}$  column of  $B$ , then

$$AB = A(\mathbf{b}_1 \ \mathbf{b}_2 \ \dots \ \mathbf{b}_p) = (A\mathbf{b}_1 \ A\mathbf{b}_2 \ \dots \ A\mathbf{b}_p), \quad (1.11)$$

indicating that the  $k^{\text{th}}$  column of their matrix product is  $A\mathbf{b}_k$ .

There are two important special matrices. The first is the *zero matrix*, all of whose entries are 0. We use  $\mathbf{O}_{m \times n}$  to denote the  $m \times n$  zero matrix, often written as just  $\mathbf{O}$  if the size is clear from the context. The zero matrix is the additive unit, so  $A + \mathbf{O} = A = \mathbf{O} + A$  when  $\mathbf{O}$  has the same size as  $A$ . In particular, we will use a bold face  $\mathbf{0}$  to denote a column vector with all zero entries, i.e.,  $\mathbf{O}_{1 \times n}$ .

The role of the multiplicative unit is played by the square *identity matrix*

$$I = I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

of size  $n \times n$ . The entries along the *main diagonal* — which runs from top left to bottom right — are equal to 1, while the *off-diagonal* entries are all 0. As you can check, if  $A$  is

---

<sup>†</sup> A much simpler — but more abstract proof can be found in Exercise 7.1.45.

---

### Basic Matrix Arithmetic

---

Matrix Addition:	Commutativity	$A + B = B + A$
	Associativity	$(A + B) + C = A + (B + C)$
	Zero Matrix	$A + O = A = O + A$
	Additive Inverse	$A + (-A) = O, \quad -A = (-1)A$
Scalar Multiplication:	Associativity	$c(dA) = (cd)A$
	Distributivity	$c(A + B) = (cA) + (cB)$ $(c + d)A = (cA) + (dA)$
	Unit Scalar	$1A = A$
	Zero Scalar	$0A = O$
Matrix Multiplication:	Associativity	$(AB)C = A(BC)$
	Distributivity	$A(B + C) = AB + AC,$ $(A + B)C = AC + BC,$
	Compatibility	$c(AB) = (cA)B = A(cB)$
	Identity Matrix	$AI = A = IA$
	Zero Matrix	$AO = O, \quad OA = O$

---

any  $m \times n$  matrix, then  $I_m A = A = A I_n$ . We will sometimes write the preceding equation as just  $IA = A = AI$ , since each matrix product is well-defined for exactly one size of identity matrix.

The identity matrix is a particular example of a *diagonal matrix*. In general, a square matrix  $A$  is diagonal if all its off-diagonal entries are zero:  $a_{ij} = 0$  for all  $i \neq j$ . We will sometimes write  $D = \text{diag}(c_1, \dots, c_n)$  for the  $n \times n$  diagonal matrix with diagonal entries  $d_{ii} = c_i$ . Thus,  $\text{diag}(1, 3, 0)$  refers to the diagonal matrix  $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ , while the  $4 \times 4$  identity matrix can be written as

$$I_4 = \text{diag}(1, 1, 1, 1) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Let us conclude this section by summarizing the basic properties of matrix arithmetic. In the accompanying table,  $A, B, C$  are matrices;  $c, d$  are scalars;  $O$  is a zero matrix; and  $I$  is an identity matrix. All matrices are assumed to have the correct sizes so that the indicated operations are defined.

---

## Exercises

- 1.2.6. (a) Write down the  $5 \times 5$  identity and zero matrices. (b) Write down their sum and their product. Does the order of multiplication matter?

1.2.7. Consider the matrices  $A = \begin{pmatrix} 1 & -1 & 3 \\ -1 & 4 & -2 \\ 3 & 0 & 6 \end{pmatrix}$ ,  $B = \begin{pmatrix} -6 & 0 & 3 \\ 4 & 2 & -1 \end{pmatrix}$ ,  $C = \begin{pmatrix} 2 & -3 \\ -3 & -4 \\ 1 & 2 \end{pmatrix}$ .

Compute the indicated combinations where possible. (a)  $3A - B$ , (b)  $AB$ , (c)  $BA$ , (d)  $(A+B)C$ , (e)  $A+BC$ , (f)  $A+2CB$ , (g)  $BCB - I$ , (h)  $A^2 - 3A + I$ , (i)  $(B - I)(C + I)$ .

1.2.8. Which of the following pairs of matrices commute under matrix multiplication?

(a)  $\begin{pmatrix} 1 & 2 \\ -2 & 1 \end{pmatrix}$ ,  $\begin{pmatrix} 2 & 3 \\ 5 & 0 \end{pmatrix}$ , (b)  $\begin{pmatrix} 3 & -1 \\ 0 & 2 \\ 1 & 4 \end{pmatrix}$ ,  $\begin{pmatrix} 4 & 2 & -2 \\ 5 & 2 & 4 \end{pmatrix}$ , (c)  $\begin{pmatrix} 3 & 0 & -1 \\ -2 & -1 & 2 \\ 2 & 0 & 0 \end{pmatrix}$ ,  $\begin{pmatrix} 2 & 0 & -1 \\ 1 & 1 & -1 \\ 2 & 0 & -1 \end{pmatrix}$ .

1.2.9. List the diagonal entries of  $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix}$ .

1.2.10. Write out the following diagonal matrices: (a)  $\text{diag}(1, 0, -1)$ , (b)  $\text{diag}(2, -2, 3, -3)$ .

1.2.11. *True or false:* (a) The sum of two diagonal matrices of the same size is a diagonal matrix. (b) The product is also diagonal.

◇ 1.2.12. (a) Show that if  $D = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$  is a  $2 \times 2$  diagonal matrix with  $a \neq b$ , then the only matrices that commute (under matrix multiplication) with  $D$  are other  $2 \times 2$  diagonal matrices. (b) What if  $a = b$ ? (c) Find all matrices that commute with  $D = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}$ , where  $a, b, c$  are all different. (d) Answer the same question for the case when  $a \neq b = c$ . (e) Prove that a matrix  $A$  commutes with an  $n \times n$  diagonal matrix  $D$  with all *distinct* diagonal entries if and only if  $A$  is a diagonal matrix.

1.2.13. Show that the matrix products  $AB$  and  $BA$  have the same size if and only if  $A$  and  $B$  are square matrices of the same size.

1.2.14. Find all matrices  $B$  that commute (under matrix multiplication) with  $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ .

1.2.15. (a) Show that, if  $A, B$  are commuting square matrices, then  $(A + B)^2 = A^2 + 2AB + B^2$ . (b) Find a pair of  $2 \times 2$  matrices  $A, B$  such that  $(A + B)^2 \neq A^2 + 2AB + B^2$ .

1.2.16. Show that if the matrices  $A$  and  $B$  commute, then they necessarily are both square and the same size.

1.2.17. Let  $A$  be an  $m \times n$  matrix. What are the permissible sizes for the zero matrices appearing in the identities  $A\mathbf{0} = \mathbf{0}$  and  $\mathbf{0}A = \mathbf{0}$ ?

1.2.18. Let  $A$  be an  $m \times n$  matrix and let  $c$  be a scalar. Show that if  $cA = \mathbf{0}$ , then either  $c = 0$  or  $A = \mathbf{0}$ .

1.2.19. *True or false:* If  $AB = \mathbf{0}$  then either  $A = \mathbf{0}$  or  $B = \mathbf{0}$ .

1.2.20. *True or false:* If  $A, B$  are square matrices of the same size, then

$$A^2 - B^2 = (A + B)(A - B).$$

1.2.21. Prove that  $A\mathbf{v} = \mathbf{0}$  for every vector  $\mathbf{v}$  (with the appropriate number of entries) if and only if  $A = \mathbf{0}$  is the zero matrix. *Hint:* If you are stuck, first try to find a proof when  $A$  is a small matrix, e.g., of size  $2 \times 2$ .

1.2.22. (a) Under what conditions is the square  $A^2$  of a matrix defined? (b) Show that  $A$  and  $A^2$  commute. (c) How many matrix multiplications are needed to compute  $A^n$ ?

1.2.23. Find a nonzero matrix  $A \neq \mathbf{0}$  such that  $A^2 = \mathbf{0}$ .

◇ 1.2.24. Let  $A$  have a row all of whose entries are zero. (a) Explain why the product  $AB$  also has a zero row. (b) Find an example where  $BA$  does not have a zero row.

1.2.25. (a) Find all solutions  $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$  to the matrix equation  $AX = I$  when

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 1 \end{pmatrix}. \quad (b) \text{ Find all solutions to } XA = I. \text{ Are they the same?}$$

1.2.26. (a) Find all solutions  $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$  to the matrix equation  $AX = B$  when

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 3 \end{pmatrix} \text{ and } B = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}. \quad (b) \text{ Find all solutions to } XA = B. \text{ Are they the same?}$$

1.2.27. (a) Find all solutions  $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$  to the matrix equation  $AX = XB$  when

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 0 & 1 \\ 3 & 0 \end{pmatrix}. \quad (b) \text{ Can you find a pair of nonzero matrices } A \neq B \text{ such}$$

that the matrix equation  $AX = XB$  has a nonzero solution  $X \neq O$ ?

1.2.28. Let  $A$  be a matrix and  $c$  a scalar. Find all solutions to the matrix equation  $cA = I$ .

◇ 1.2.29. Let  $\mathbf{e}$  be the  $1 \times m$  row vector all of whose entries are equal to 1. (a) Show that if  $A$  is an  $m \times n$  matrix, then the  $i^{\text{th}}$  entry of the product  $\mathbf{v} = \mathbf{e}A$  is the  $j^{\text{th}}$  column sum of  $A$ , meaning the sum of all the entries in its  $j^{\text{th}}$  row. (b) Let  $W$  denote the  $m \times m$  matrix whose diagonal entries are equal to  $\frac{1-m}{m}$  and whose off-diagonal entries are all equal to  $\frac{1}{m}$ . Prove that the column sums of  $B = WA$  are all zero. (c) Check both results when  $A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 3 \\ -4 & 5 & -1 \end{pmatrix}$ . **Remark.** If the rows of  $A$  represent experimental data values, then the entries of  $\frac{1}{m}\mathbf{e}A$  represent the means or averages of the data values, while  $B = WA$  corresponds to data that has been normalized to have mean 0; see Section 8.8.

◇ 1.2.30. The *commutator* of two matrices  $A, B$ , is defined to be the matrix

$$C = [A, B] = AB - BA. \quad (1.12)$$

(a) Explain why  $[A, B]$  is defined if and only if  $A$  and  $B$  are square matrices of the same size. (b) Show that  $A$  and  $B$  commute under matrix multiplication if and only if  $[A, B] = O$ .

(c) Compute the commutator of the following matrices:

$$(i) \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ -2 & 0 \end{pmatrix}; \quad (ii) \begin{pmatrix} -1 & 3 \\ 3 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 7 \\ 7 & 1 \end{pmatrix}; \quad (iii) \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix};$$

(d) Prove that the commutator is (i) *Bilinear*:  $[cA + dB, C] = c[A, C] + d[B, C]$  and  $[A, cB + dC] = c[A, B] + d[A, C]$  for any scalars  $c, d$ ; (ii) *Skew-symmetric*:  $[A, B] = -[B, A]$ ; (iii) satisfies the the *Jacobi identity*:

$$[[A, B], C] + [[C, A], B] + [[B, C], A] = O,$$

for any square matrices  $A, B, C$  of the same size.

**Remark.** The commutator plays a very important role in geometry, symmetry, and quantum mechanics. See Section 10.4 as well as [54, 60, 93] for further developments.

◇ 1.2.31. The *trace* of a  $n \times n$  matrix  $A \in \mathcal{M}_{n \times n}$  is defined to be the sum of its diagonal entries:

$$\text{tr } A = a_{11} + a_{22} + \cdots + a_{nn}. \quad (a) \text{ Compute the trace of (i) } \begin{pmatrix} 1 & -1 \\ 2 & 3 \end{pmatrix}, \text{ (ii) } \begin{pmatrix} 1 & 3 & 2 \\ -1 & 0 & 1 \\ -4 & 3 & -1 \end{pmatrix}.$$

(b) Prove that  $\text{tr}(A + B) = \text{tr } A + \text{tr } B$ . (c) Prove that  $\text{tr}(AB) = \text{tr}(BA)$ . (d) Prove that the commutator matrix  $C = AB - BA$  has zero trace:  $\text{tr } C = 0$ . (e) Is part (c) valid if  $A$  has size  $m \times n$  and  $B$  has size  $n \times m$ ? (f) Prove that  $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$ . On the other hand, find an example where  $\text{tr}(ABC) \neq \text{tr}(ACB)$ .

◇ 1.2.32. Prove that matrix multiplication is associative:  $A(BC) = (AB)C$  when defined.

◇ 1.2.33. Justify the following alternative formula for multiplying a matrix  $A$  and a column vector  $\mathbf{x}$ :

$$A\mathbf{x} = x_1 \mathbf{c}_1 + x_2 \mathbf{c}_2 + \cdots + x_n \mathbf{c}_n, \quad (1.13)$$

where  $\mathbf{c}_1, \dots, \mathbf{c}_n$  are the columns of  $A$  and  $x_1, \dots, x_n$  the entries of  $\mathbf{x}$ .

◇ 1.2.34. The basic definition of matrix multiplication  $AB$  tells us to multiply rows of  $A$  by columns of  $B$ . Remarkably, if you suitably interpret the operation, you can also compute  $AB$  by multiplying columns of  $A$  by rows of  $B$ ! Suppose  $A$  is an  $m \times n$  matrix with columns  $\mathbf{c}_1, \dots, \mathbf{c}_n$ . Suppose  $B$  is an  $n \times p$  matrix with rows  $\mathbf{r}_1, \dots, \mathbf{r}_n$ . Then we claim that

$$AB = \mathbf{c}_1 \mathbf{r}_1 + \mathbf{c}_2 \mathbf{r}_2 + \cdots + \mathbf{c}_n \mathbf{r}_n, \quad (1.14)$$

where each summand is a matrix of size  $m \times p$ . (a) Verify that the particular case

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} (0 \ -1) + \begin{pmatrix} 2 \\ 4 \end{pmatrix} (2 \ 3) = \begin{pmatrix} 0 & -1 \\ 0 & -3 \end{pmatrix} + \begin{pmatrix} 4 & 6 \\ 8 & 12 \end{pmatrix} = \begin{pmatrix} 4 & 5 \\ 8 & 9 \end{pmatrix}$$

agrees with the usual method for computing the matrix product. (b) Use this method to compute the matrix products (i)  $\begin{pmatrix} -2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 1 & 0 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 1 & -2 & 0 \\ -3 & -1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 5 \\ -3 & 0 \\ 1 & -1 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 3 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & -5 \end{pmatrix} \begin{pmatrix} 2 & 3 & 0 \\ 3 & -1 & 4 \\ 0 & 4 & 1 \end{pmatrix}$ , and verify that you get the same answer as that

obtained by the traditional method. (c) Explain why (1.13) is a special case of (1.14). (d) Prove that (1.14) gives the correct formula for the matrix product.

◇ 1.2.35. *Matrix polynomials.* Let  $p(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0$  be a polynomial function. If  $A$  is a square matrix, we define the corresponding *matrix polynomial*  $p(A) = c_n A^n + c_{n-1} A^{n-1} + \cdots + c_1 A + c_0 \mathbf{I}$ ; the constant term becomes a scalar multiple of the identity matrix. For instance, if  $p(x) = x^2 - 2x + 3$ , then  $p(A) = A^2 - 2A + 3\mathbf{I}$ . (a) Write out the matrix polynomials  $p(A), q(A)$  when  $p(x) = x^3 - 3x + 2$ ,  $q(x) = 2x^2 + 1$ . (b) Evaluate  $p(A)$  and  $q(A)$  when  $A = \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix}$ . (c) Show that the matrix product  $p(A)q(A)$  is the matrix polynomial corresponding to the product polynomial  $r(x) = p(x)q(x)$ . (d) *True or false:* If  $B = p(A)$  and  $C = q(A)$ , then  $BC = CB$ . Check your answer in the particular case of part (b).

◇ 1.2.36. A *block matrix* has the form  $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$  in which  $A, B, C, D$  are matrices with respective sizes  $i \times k, i \times l, j \times k, j \times l$ . (a) What is the size of  $M$ ? (b) Write out the

block matrix  $M$  when  $A = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ ,  $B = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ ,  $C = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$ ,  $D = \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ 1 & -1 \end{pmatrix}$ .

(c) Show that if  $N = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$  is a block matrix whose blocks have the same size as those of  $M$ , then  $M + N = \begin{pmatrix} A+P & B+Q \\ C+R & D+S \end{pmatrix}$ , i.e., matrix addition can be done in blocks.

(d) Show that if  $P = \begin{pmatrix} X & Y \\ Z & W \end{pmatrix}$  has blocks of a compatible size, the matrix product is

$MP = \begin{pmatrix} AX + BZ & AY + BW \\ CX + DZ & CY + DW \end{pmatrix}$ . Explain what “compatible” means. (e) Write down a compatible block matrix  $P$  for the matrix  $M$  in part (b). Then validate the block matrix product identity of part (d) for your chosen matrices.

- ◇ 1.2.37. The matrix  $S$  is said to be a *square root* of the matrix  $A$  if  $S^2 = A$ . (a) Show that  $S = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix}$  is a square root of the matrix  $A = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$ . Can you find another square root of  $A$ ? (b) Explain why only square matrices can have a square root. (c) Find all real square roots of the  $2 \times 2$  identity matrix  $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ . (d) Does  $-I = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$  have a real square root?

### 1.3 Gaussian Elimination — Regular Case

With the basic matrix arithmetic operations in hand, let us now return to our primary task. The goal is to develop a systematic method for solving linear systems of equations. While we could continue to work directly with the equations, matrices provide a convenient alternative that begins by merely shortening the amount of writing, but ultimately leads to profound insight into the structure of linear systems and their solutions.

We begin by replacing the system (1.7) by its matrix constituents. It is convenient to ignore the vector of unknowns, and form the *augmented matrix*

$$M = (A \mid \mathbf{b}) = \left( \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array} \right), \quad (1.15)$$

which is an  $m \times (n + 1)$  matrix obtained by tacking the right-hand side vector onto the original coefficient matrix. The extra vertical line is included just to remind us that the last column of this matrix plays a special role. For example, the augmented matrix for the system (1.1), i.e.,

$$\begin{aligned} x + 2y + z &= 2, \\ 2x + 6y + z &= 7, \\ x + y + 4z &= 3, \end{aligned} \quad \text{is} \quad M = \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array} \right). \quad (1.16)$$

Note that one can immediately recover the equations in the original linear system from the augmented matrix. Since operations on equations also affect their right-hand sides, keeping track of everything is most easily done through the augmented matrix.

For the time being, we will concentrate our efforts on linear systems that have the same number,  $n$ , of equations as unknowns. The associated coefficient matrix  $A$  is square, of size  $n \times n$ . The corresponding augmented matrix  $M = (A \mid \mathbf{b})$  then has size  $n \times (n + 1)$ .

The matrix operation that assumes the role of Linear System Operation #1 is:

#### Elementary Row Operation #1:

Add a scalar multiple of one row of the augmented matrix to another row.

For example, if we add  $-2$  times the first row of the augmented matrix (1.16) to the second row, the result is the row vector

$$-2(1 \ 2 \ 1 \ 2) + (2 \ 6 \ 1 \ 7) = (0 \ 2 \ -1 \ 3).$$

The result can be recognized as the second row of the modified augmented matrix

$$\left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 1 & 1 & 4 & 3 \end{array} \right) \quad (1.17)$$

that corresponds to the first equivalent system (1.2). When elementary row operation #1 is performed, it is critical that the result replaces the row being added to — *not* the row being multiplied by the scalar. Notice that the elimination of a variable in an equation — in this case, the first variable in the second equation — amounts to making its entry in the coefficient matrix equal to zero.

We shall call the (1, 1) entry of the coefficient matrix the *first pivot*. The precise definition of pivot will become clear as we continue; the one key requirement is that *a pivot must always be nonzero*. Eliminating the first variable  $x$  from the second and third equations amounts to making all the matrix entries in the column below the pivot equal to zero. We have already done this with the (2, 1) entry in (1.17). To make the (3, 1) entry equal to zero, we subtract (that is, add  $-1$  times) the first row from the last row. The resulting augmented matrix is

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

which corresponds to the system (1.3). The *second pivot* is the (2, 2) entry of this matrix, which is 2, and is the coefficient of the second variable in the second equation. Again, the pivot must be nonzero. We use the elementary row operation of adding  $\frac{1}{2}$  of the second row to the third row to make the entry below the second pivot equal to 0; the result is the augmented matrix

$$N = \left( \begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 0 & \frac{5}{2} & \frac{5}{2} \end{array} \right)$$

that corresponds to the triangular system (1.4). We write the final augmented matrix as

$$N = (U | \mathbf{c}), \quad \text{where} \quad U = \left( \begin{array}{ccc} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{array} \right), \quad \mathbf{c} = \left( \begin{array}{c} 2 \\ 3 \\ \frac{5}{2} \end{array} \right).$$

The corresponding linear system has vector form

$$U \mathbf{x} = \mathbf{c}. \quad (1.18)$$

Its coefficient matrix  $U$  is *upper triangular*, which means that all its entries below the main diagonal are zero:  $u_{ij} = 0$  whenever  $i > j$ . The three nonzero entries on its diagonal, 1, 2,  $\frac{5}{2}$ , including the last one in the (3, 3) slot, are the three pivots. Once the system has been reduced to triangular form (1.18), we can easily solve it by Back Substitution.

The preceding algorithm for solving a linear system of  $n$  equations in  $n$  unknowns is known as *regular Gaussian Elimination*. A square matrix  $A$  will be called *regular*<sup>†</sup> if the algorithm successfully reduces it to upper triangular form  $U$  with all non-zero pivots on the diagonal. In other words, for regular matrices, as the algorithm proceeds, each successive pivot appearing on the diagonal must be nonzero; otherwise, the matrix is not regular. We then use the pivot row to make all the entries lying in the column below the pivot equal to zero through elementary row operations. The solution is found by applying Back Substitution to the resulting triangular system.

---

<sup>†</sup> Strangely, there is no commonly accepted term to describe this kind of matrix. For lack of a better alternative, we propose to use the adjective “regular” in the sequel.

---

 Gaussian Elimination — Regular Case
 

---



---

```

start
  for  $j = 1$  to  $n$ 
    if  $m_{jj} = 0$ , stop; print “ $A$  is not regular”
    else for  $i = j + 1$  to  $n$ 
      set  $l_{ij} = m_{ij}/m_{jj}$ 
      add  $-l_{ij}$  times row  $j$  of  $M$  to row  $i$  of  $M$ 
    next  $i$ 
  next  $j$ 
end

```

---

Let us state this algorithm in the form of a program, written in a general “pseudocode” that can be easily translated into any specific language, e.g., C++, FORTRAN, JAVA, MAPLE, MATHEMATICA, MATLAB. In accordance with the usual programming convention, the same letter  $M = (m_{ij})$  will be used to denote the current augmented matrix at each stage in the computation, keeping in mind that its entries will change as the algorithm progresses. We initialize  $M = (A | \mathbf{b})$ . The final output of the program, assuming  $A$  is regular, is the augmented matrix  $M = (U | \mathbf{c})$ , where  $U$  is the upper triangular matrix whose diagonal entries are the pivots, while  $\mathbf{c}$  is the resulting vector of right-hand sides in the triangular system  $U \mathbf{x} = \mathbf{c}$ .

For completeness, let us include the pseudocode program for Back Substitution. The input to this program is the upper triangular matrix  $U$  and the right-hand side vector  $\mathbf{c}$  that results from the Gaussian Elimination pseudocode program, which produces  $M = (U | \mathbf{c})$ . The output of the Back Substitution program is the solution vector  $\mathbf{x}$  to the triangular system  $U \mathbf{x} = \mathbf{c}$ , which is the *same* as the solution to the original linear system  $A \mathbf{x} = \mathbf{b}$ .

---

 Back Substitution
 

---



---

```

start
  set  $x_n = c_n/u_{nn}$ 
  for  $i = n - 1$  to  $1$  with increment  $-1$ 
    set  $x_i = \frac{1}{u_{ii}} \left( c_i - \sum_{j=1}^{i+1} u_{ij} x_j \right)$ 
  next  $j$ 
end

```

---



---

 Exercises
 

---

- 1.3.1. Solve the following linear systems by Gaussian Elimination. (a)  $\begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 7 \\ 3 \end{pmatrix}$ ,

$$(b) \begin{pmatrix} 6 & -1 \\ 3 & -2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & 1 & 2 \\ -1 & 3 & 3 \\ 4 & -3 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \\ 7 \end{pmatrix},$$

$$(d) \begin{pmatrix} 5 & 3 & -1 \\ 3 & 2 & -1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 9 \\ 5 \\ -1 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 & 1 & -1 \\ 2 & -1 & 3 \\ -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} 0 \\ 3 \\ 5 \end{pmatrix},$$

$$(f) \begin{pmatrix} -1 & 1 & 1 & 0 \\ 2 & -1 & 0 & 1 \\ 1 & 0 & 2 & 3 \\ 0 & 1 & -1 & -2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad (g) \begin{pmatrix} 2 & -3 & 1 & 1 \\ 1 & -1 & -2 & -1 \\ 3 & -2 & 1 & 2 \\ 1 & 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 5 \\ 3 \end{pmatrix}.$$

1.3.2. Write out the augmented matrix for the following linear systems. Then solve the system by first applying elementary row operations of type #1 to place the augmented matrix in upper triangular form, followed by Back Substitution.

$$(a) \begin{array}{l} x_1 + 7x_2 = 4, \\ -2x_1 - 9x_2 = 2. \end{array} \quad (b) \begin{array}{l} 3z - 5w = -1, \\ 2z + w = 8. \end{array} \quad (c) \begin{array}{l} x - 2y + z = 0, \\ 2y - 8z = 8, \\ -4x + 5y + 9z = -9. \end{array}$$

$$(d) \begin{array}{l} p + 4q - 2r = 1, \\ -2p - 3r = -7, \\ 3p - 2q + 2r = -1. \end{array} \quad (e) \begin{array}{l} x_1 - 2x_3 = -1, \\ x_2 - x_4 = 2, \\ -3x_2 + 2x_3 = 0, \\ -4x_1 + 7x_4 = -5. \end{array} \quad (f) \begin{array}{l} -x + 3y - z + w = -2, \\ x - y + 3z - w = 0, \\ y - z + 4w = 7, \\ 4x - y + z = 5. \end{array}$$

1.3.3. For each of the following augmented matrices write out the corresponding linear system of equations. Solve the system by applying Gaussian Elimination to the augmented matrix.

$$(a) \begin{pmatrix} 3 & 2 & & 2 \\ -4 & -3 & & -1 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 2 & 0 & -3 \\ -1 & 2 & 1 & -6 \\ -2 & 0 & -3 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 \\ 0 & -1 & 2 & -1 & 1 \\ 0 & 0 & -1 & 2 & 0 \end{pmatrix}.$$

1.3.4. Which of the following matrices are regular?

$$(a) \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & -1 \\ 3 & -2 \end{pmatrix},$$

$$(c) \begin{pmatrix} 3 & -2 & 1 \\ -1 & 4 & -3 \\ 3 & -2 & 5 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & -2 & 3 \\ -2 & 4 & -1 \\ 3 & -1 & 2 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 & 3 & -3 & 0 \\ -1 & 0 & -1 & 2 \\ 3 & 3 & -6 & 1 \\ 2 & 3 & -3 & 5 \end{pmatrix}.$$

1.3.5. The techniques that are developed for solving linear systems are also applicable to systems with complex coefficients, whose solutions may also be complex. Use Gaussian Elimination to solve the following complex linear systems.

$$(a) \begin{array}{l} -i x_1 + (1+i)x_2 = -1, \\ (1-i)x_1 + x_2 = -3i. \end{array} \quad (b) \begin{array}{l} ix + (1-i)z = 2i, \\ 2iy + (1+i)z = 2, \\ -x + 2iy + iz = 1 - 2i. \end{array}$$

$$(c) \begin{array}{l} (1-i)x + 2y = i, \\ -ix + (1+i)y = -1. \end{array} \quad (d) \begin{array}{l} (1+i)x + iy + (2+2i)z = 0, \\ (1-i)x + 2y + iz = 0, \\ (3-3i)x + iy + (3-11i)z = 6. \end{array}$$

1.3.6. (a) Write down an example of a system of 5 linear equations in 5 unknowns with regular diagonal coefficient matrix. (b) Solve your system. (c) Explain why solving a system whose coefficient matrix is diagonal is very easy.

1.3.7. Find the equation of the parabola  $y = ax^2 + bx + c$  that goes through the points  $(1, 6), (2, 4)$ , and  $(3, 0)$ .

◇ 1.3.8. A linear system is called *homogeneous* if all the right-hand sides are zero, and so takes the matrix form  $A\mathbf{x} = \mathbf{0}$ . Explain why the solution to a homogeneous system with regular coefficient matrix is  $\mathbf{x} = \mathbf{0}$ .

1.3.9. Under what conditions do two  $2 \times 2$  upper triangular matrices commute?

1.3.10. A matrix is called *lower triangular* if all entries above the diagonal are zero. Show that a matrix is both lower and upper triangular if and only if it is a diagonal matrix.

◇ 1.3.11. A square matrix is called *strictly lower (upper) triangular* if all entries on or above (below) the main diagonal are 0. (a) Prove that every square matrix can be uniquely written as a sum  $A = L + D + U$ , with  $L$  strictly lower triangular,  $D$  diagonal, and  $U$

strictly upper triangular. (b) Decompose  $A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & 0 & 5 \end{pmatrix}$  in this manner.

◇ 1.3.12. A square matrix  $N$  is called *nilpotent* if  $N^k = \mathbf{0}$  for some  $k \geq 1$ .

(a) Show that  $N = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$  is nilpotent. (b) Show that every strictly upper triangular matrix, as defined in Exercise 1.3.11, is nilpotent. (c) Find a nilpotent matrix which is neither lower nor upper triangular.

◇ 1.3.13. A square matrix  $W$  is called *unipotent* if  $N = W - \mathbf{I}$  is nilpotent, as in Exercise 1.3.12, so  $(W - \mathbf{I})^k = \mathbf{0}$  for some  $k \geq 1$ . (a) Show that every lower or upper triangular matrix is unipotent if and only if it is unitriangular, meaning its diagonal entries are all equal to 1.

(b) Find a unipotent matrix which is neither lower nor upper triangular.

1.3.14. A square matrix  $P$  is called *idempotent* if  $P^2 = P$ . (a) Find all  $2 \times 2$  idempotent upper triangular matrices. (b) Find all  $2 \times 2$  idempotent matrices.

## Elementary Matrices

A key observation is that elementary row operations can, in fact, be realized by matrix multiplication. To this end, we introduce the first type of “elementary matrix”. (Later we will meet two other types of elementary matrix, corresponding to the other two kinds of elementary row operation.)

**Definition 1.1.** The *elementary matrix* associated with an elementary row operation for  $m$ -rowed matrices is the  $m \times m$  matrix obtained by applying the row operation to the  $m \times m$  identity matrix  $\mathbf{I}_m$ .

For example, applying the elementary row operation that adds  $-2$  times the first row to the second row of the  $3 \times 3$  identity matrix  $\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$  results in the corresponding elementary matrix  $E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ . We claim that, if  $A$  is any  $3$ -rowed matrix, then

multiplying  $E_1 A$  has the same effect as the given elementary row operation. For example,

$$\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 1 & 1 & 4 \end{pmatrix},$$

which you may recognize as the first elementary row operation we used to solve our

illustrative example. If we set

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}, \quad (1.19)$$

then multiplication by  $E_1$  will subtract twice the first row from the second row, multiplication by  $E_2$  will subtract the first row from the third row, and multiplication by  $E_3$  will add  $\frac{1}{2}$  the second row to the third row — precisely the row operations used to place our original system in triangular form. Therefore, performing them in the correct order, we conclude that when

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}, \quad \text{then} \quad E_3 E_2 E_1 A = U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}. \quad (1.20)$$

The reader is urged to check this by directly multiplying the indicated matrices. Keep in mind that the associative property of matrix multiplication allows us to compute the above matrix product in any convenient order:

$$E_3 E_2 E_1 A = E_3 (E_2 (E_1 A)) = ((E_3 E_2) E_1) A = (E_3 (E_2 E_1)) A = (E_3 E_2) (E_1 A) = \dots,$$

making sure that the overall left to right order of the matrices is maintained, since the matrix products are usually *not* commutative.

In general, then, an  $m \times m$  *elementary matrix*  $E$  of the first type will have all 1's on the diagonal, one nonzero entry  $c$  in some off-diagonal position  $(i, j)$ , with  $i \neq j$ , and all other entries equal to zero. If  $A$  is any  $m \times n$  matrix, then the matrix product  $EA$  is equal to the matrix obtained from  $A$  by the elementary row operation adding  $c$  times row  $j$  to row  $i$ . (Note that the order of  $i$  and  $j$  is reversed.)

To undo the operation of adding  $c$  times row  $j$  to row  $i$ , we must perform the inverse row operation that subtracts  $c$  (or, equivalently, adds  $-c$ ) times row  $j$  from row  $i$ . The corresponding *inverse elementary matrix* again has 1's along the diagonal and  $-c$  in the  $(i, j)$  slot. Let us denote the inverses of the particular elementary matrices (1.19) by  $L_i$ , so that, according to our general rule,

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}. \quad (1.21)$$

Note that the products

$$L_1 E_1 = L_2 E_2 = L_3 E_3 = I \quad (1.22)$$

yield the  $3 \times 3$  identity matrix, reflecting the fact that the matrices represent mutually inverse row operations. (A more thorough discussion of matrix inverses will be postponed until Section 1.5.)

The product of the latter three elementary matrices (1.21) is equal to

$$L = L_1 L_2 L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix}. \quad (1.23)$$

The matrix  $L$  is called a *lower unitriangular* matrix, where “lower triangular” means that all the entries above the main diagonal are 0, while “uni-”, which is short for “unipotent”

as defined in Exercise 1.3.13, imposes the requirement that all the entries on the diagonal are equal to 1. Observe that the entries of  $L$  below the diagonal are the same as the corresponding nonzero entries in the  $L_i$ . This is a general fact that holds when the lower triangular elementary matrices are multiplied in the correct order. More generally, the following elementary consequence of the laws of matrix multiplication will be used extensively.

**Lemma 1.2.** If  $L$  and  $\widehat{L}$  are lower triangular matrices of the same size, so is their product  $L\widehat{L}$ . If they are both lower unitriangular, so is their product. Similarly, if  $U, \widehat{U}$  are upper (uni)triangular matrices, so is their product  $U\widehat{U}$ .

### The $LU$ Factorization

We have almost arrived at our first important result. Let us compute the product of the matrices  $L$  and  $U$  in (1.20), (1.23). Using associativity of matrix multiplication, equations (1.22), and the basic property of the identity matrix  $I$ , we conclude that

$$\begin{aligned} LU &= (L_1 L_2 L_3)(E_3 E_2 E_1 A) = L_1 L_2 (L_3 E_3) E_2 E_1 A = L_1 L_2 I E_2 E_1 A \\ &= L_1 (L_2 E_2) E_1 A = L_1 I E_1 A = (L_1 E_1) A = I A = A. \end{aligned}$$

In other words, we have *factored* the coefficient matrix  $A = LU$  into a product of a lower unitriangular matrix  $L$  and an upper triangular matrix  $U$  with the nonzero pivots on its main diagonal. By similar reasoning, the same holds true for any regular square matrix.

**Theorem 1.3.** A matrix  $A$  is regular if and only if it can be factored

$$A = LU, \quad (1.24)$$

where  $L$  is a lower unitriangular matrix, having all 1's on the diagonal, and  $U$  is upper triangular with nonzero diagonal entries, which are the pivots of  $A$ . The nonzero off-diagonal entries  $l_{ij}$  for  $i > j$  appearing in  $L$  prescribe the elementary row operations that bring  $A$  into upper triangular form; namely, one subtracts  $l_{ij}$  times row  $j$  from row  $i$  at the appropriate step of the Gaussian Elimination process.

In practice, to find the  $LU$  factorization of a square matrix  $A$ , one applies the regular Gaussian Elimination algorithm to reduce  $A$  to its upper triangular form  $U$ . The entries of  $L$  can be filled in during the course of the calculation with the negatives of the multiples used in the elementary row operations. If the algorithm fails to be completed, which happens whenever zero appears in any diagonal pivot position, then the original matrix is *not* regular, and does *not* have an  $LU$  factorization.

**Example 1.4.** Let us compute the  $LU$  factorization of the matrix  $A = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 5 & 2 \\ 2 & -2 & 0 \end{pmatrix}$ .

Applying the Gaussian Elimination algorithm, we begin by adding  $-2$  times the first row to the second row, and then adding  $-1$  times the first row to the third. The result is the

matrix  $\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & -3 & -1 \end{pmatrix}$ . The next step adds the second row to the third row, leading to the upper triangular matrix  $U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}$ , whose diagonal entries are the pivots. The

corresponding lower triangular matrix is  $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$ ; its entries lying below the

main diagonal are the *negatives* of the multiples we used during the elimination procedure. For instance, the  $(2, 1)$  entry indicates that we added  $-2$  times the first row to the second row, and so on. The reader might wish to verify the resulting factorization

$$\begin{pmatrix} 2 & 1 & 1 \\ 4 & 5 & 2 \\ 2 & -2 & 0 \end{pmatrix} = A = LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

## Exercises

1.3.15. What elementary row operations do the following matrices represent? What size matrices do they apply to?

$$(a) \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}, (b) \begin{pmatrix} 1 & 0 \\ 7 & 1 \end{pmatrix}, (c) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -5 \\ 0 & 0 & 1 \end{pmatrix}, (d) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix}, (e) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

1.3.16. Write down the elementary matrix corresponding to the following row operations on  $4 \times 4$  matrices: (a) Add the third row to the fourth row. (b) Subtract the fourth row from the third row. (c) Add 3 times the last row to the first row. (d) Subtract twice the second row from the fourth row.

1.3.17. Compute the product  $L_3 L_2 L_1$  of the elementary matrices (1.21). Compare your answer with (1.23).

1.3.18. Determine the product  $E_3 E_2 E_1$  of the elementary matrices in (1.19). Is this the same as the product  $E_1 E_2 E_3$ ? Which is easier to predict?

1.3.19. (a) Explain, using their interpretation as elementary row operations, why elementary matrices do not generally commute:  $E \tilde{E} \neq \tilde{E} E$ . (b) Which pairs of the elementary matrices listed in (1.19) commute? (c) Can you formulate a general rule that tells in advance whether two given elementary matrices commute?

1.3.20. Determine which of the following  $3 \times 3$  matrices is (i) upper triangular, (ii) upper unitriangular, (iii) lower triangular, and/or (iv) lower unitriangular:

$$(a) \begin{pmatrix} 1 & 2 & 0 \\ 0 & 3 & 2 \\ 0 & 0 & -2 \end{pmatrix} (b) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} (c) \begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 3 & 3 \end{pmatrix} (d) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -4 & 1 \end{pmatrix} (e) \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

1.3.21. Find the  $LU$  factorization of the following matrices: (a)  $\begin{pmatrix} 1 & 3 \\ -1 & 0 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} -1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & 0 & 3 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ , (e)  $\begin{pmatrix} -1 & 0 & 0 \\ 2 & -3 & 0 \\ 1 & 3 & 2 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 0 & -1 \\ 2 & 3 & 2 \\ -3 & 1 & 0 \end{pmatrix}$ , (g)  $\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 2 & -1 & -1 \\ -1 & 3 & 0 & 2 \\ 0 & -1 & 2 & 1 \end{pmatrix}$ , (h)  $\begin{pmatrix} 1 & 1 & -2 & 3 \\ -1 & 2 & 3 & 0 \\ -2 & 1 & 1 & -2 \\ 3 & 0 & 1 & 5 \end{pmatrix}$ , (i)  $\begin{pmatrix} 2 & 1 & 3 & 1 \\ 1 & 4 & 0 & 1 \\ 3 & 0 & 2 & 2 \\ 1 & 1 & 2 & 2 \end{pmatrix}$ .

1.3.22. Given the factorization  $A = \begin{pmatrix} 2 & -1 & 0 \\ -6 & 4 & -1 \\ 4 & -6 & 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 2 & -4 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{pmatrix}$ ,

explain, without computing, which elementary row operations are used to reduce  $A$  to upper triangular form. Be careful to state the order in which they should be applied. Then check the correctness of your answer by performing the elimination.

- 1.3.23. (a) Write down a  $4 \times 4$  lower unitriangular matrix whose entries below the diagonal are distinct nonzero numbers. (b) Explain which elementary row operation each entry corresponds to. (c) Indicate the order in which the elementary row operations should be performed by labeling the entries 1, 2, 3, ... .

- ◊ 1.3.24. Let  $t_1, t_2, \dots$  be distinct real numbers. Find the  $LU$  factorization of the following

*Vandermonde matrices:* (a)  $\begin{pmatrix} 1 & 1 \\ t_1 & t_2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \\ t_1^2 & t_2^2 & t_3^2 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 1 & 1 & 1 \\ t_1 & t_2 & t_3 & t_4 \\ t_1^2 & t_2^2 & t_3^2 & t_4^2 \\ t_1^3 & t_2^3 & t_3^3 & t_4^3 \end{pmatrix}$ .

Can you spot a pattern? Test your conjecture with the  $5 \times 5$  Vandermonde matrix.

- 1.3.25. Write down the explicit requirements on its entries  $a_{ij}$  for a square matrix  $A$  to be

- (a) diagonal, (b) upper triangular, (c) upper unitriangular, (d) lower triangular,  
(e) lower unitriangular.

- ◊ 1.3.26. (a) Explain why the product of two lower triangular matrices is lower triangular.

- (b) What can you say concerning the diagonal entries of the product of two lower triangular matrices? (c) Explain why the product of two lower unitriangular matrices is also lower unitriangular.

- 1.3.27. *True or false:* If  $A$  has a zero entry on its main diagonal, it is not regular.

- 1.3.28. In general, how many elementary row operations does one need to perform in order to reduce a regular  $n \times n$  matrix to upper triangular form?

- 1.3.29. Prove that if  $A$  is a regular  $2 \times 2$  matrix, then its  $LU$  factorization is unique. In other words, if  $A = LU = \hat{L}\hat{U}$  where  $L, \hat{L}$  are lower unitriangular and  $U, \hat{U}$  are upper triangular, then  $L = \hat{L}$  and  $U = \hat{U}$ . (The general case appears in Proposition 1.30.)

- ◊ 1.3.30. Prove directly that the matrix  $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$  does not have an  $LU$  factorization.

- ◊ 1.3.31. Suppose  $A$  is regular. (a) Show that the matrix obtained by multiplying each column of  $A$  by the sign of its pivot is also regular and, moreover, has all positive pivots.

- (b) Show that the matrix obtained by multiplying each row of  $A$  by the sign of its pivot is also regular and has all positive pivots.

- (c) Check these results in the particular case  $A = \begin{pmatrix} -2 & 2 & 1 \\ 1 & 0 & 1 \\ 4 & 2 & 3 \end{pmatrix}$ .

## Forward and Back Substitution

Knowing the  $LU$  factorization of a regular matrix  $A$  enables us to solve any associated linear system  $A\mathbf{x} = \mathbf{b}$  in two easy stages:

- (1) First, solve the lower triangular system

$$L\mathbf{c} = \mathbf{b} \quad (1.25)$$

for the vector  $\mathbf{c}$  by *Forward Substitution*. This is the same as Back Substitution, except one solves the equations for the variables in the direct order — from first to last. Explicitly,

$$c_1 = b_1, \quad c_i = b_i - \sum_{j=1}^{i-1} l_{ij} c_j, \quad \text{for } i = 2, 3, \dots, n, \quad (1.26)$$

noting that the previously computed values of  $c_1, \dots, c_{i-1}$  are used to determine  $c_i$ .

- (2) Second, solve the resulting upper triangular system

$$U\mathbf{x} = \mathbf{c} \quad (1.27)$$

by *Back Substitution*. The values of the unknowns

$$x_n = \frac{c_n}{u_{nn}}, \quad x_i = \frac{1}{u_{ii}} \left( c_i - \sum_{j=i+1}^n u_{ij} x_j \right), \quad \text{for } i = n-1, \dots, 2, 1, \quad (1.28)$$

are successively computed, but now in reverse order. It is worth pointing out that the requirement that each pivot be nonzero,  $u_{ii} \neq 0$ , is essential here, as otherwise we would not be able to solve for the corresponding variable  $x_i$ .

Note that the combined algorithm does indeed solve the original system, since if

$$U\mathbf{x} = \mathbf{c} \quad \text{and} \quad L\mathbf{c} = \mathbf{b}, \quad \text{then} \quad A\mathbf{x} = LU\mathbf{x} = L\mathbf{c} = \mathbf{b}.$$

**Example 1.5.** With the  $LU$  decomposition

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

found in Example 1.4, we can readily solve any linear system with the given coefficient matrix by Forward and Back Substitution. For instance, to find the solution to

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

we first solve the lower triangular system

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \quad \begin{array}{ll} a & = 1, \\ 2a + b & = 2, \\ a - b + c & = 2. \end{array}$$

The first equation says  $a = 1$ ; substituting into the second, we find  $b = 0$ ; the final equation yields  $c = 1$ . We then use Back Substitution to solve the upper triangular system

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{array}{ll} 2x + y + z & = 1, \\ 3y & = 0, \\ -z & = 1. \end{array}$$

We find  $z = -1$ , then  $y = 0$ , and then  $x = 1$ , which is indeed the solution.

Thus, once we have found the  $LU$  factorization of the coefficient matrix  $A$ , the Forward and Back Substitution processes quickly produce the solution to any system  $A\mathbf{x} = \mathbf{b}$ . Moreover, they can be straightforwardly programmed on a computer. In practice, to solve a system from scratch, it is just a matter of taste whether you work directly with the augmented matrix, or first determine the  $LU$  factorization of the coefficient matrix, and then apply Forward and Back Substitution to compute the solution.

## Exercises

- 1.3.32. Given the  $LU$  factorizations you calculated in Exercise 1.3.21, solve the associated linear systems  $A\mathbf{x} = \mathbf{b}$ , where  $\mathbf{b}$  is the column vector with all entries equal to 1.

1.3.33. In each of the following problems, find the  $A = LU$  factorization of the coefficient matrix, and then use Forward and Back Substitution to solve the corresponding linear systems  $A \mathbf{x} = \mathbf{b}_j$  for each of the indicated right-hand sides:

$$(a) A = \begin{pmatrix} -1 & 3 \\ 3 & 2 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \mathbf{b}_3 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$$

$$(b) A = \begin{pmatrix} -1 & 1 & -1 \\ 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix}.$$

$$(c) A = \begin{pmatrix} 9 & -2 & -1 \\ -6 & 1 & 1 \\ 2 & -1 & 0 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}.$$

$$(d) A = \begin{pmatrix} 2.0 & .3 & .4 \\ .3 & 4.0 & .5 \\ .4 & .5 & 6.0 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

$$(e) A = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 2 & 3 & -1 \\ -1 & 3 & 2 & 2 \\ 0 & -1 & 2 & 1 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}.$$

$$(f) A = \begin{pmatrix} 1 & -2 & 0 & 2 \\ 4 & 1 & -1 & -1 \\ -8 & -1 & 2 & 1 \\ -4 & -1 & 1 & 2 \end{pmatrix}, \mathbf{b}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{b}_2 = \begin{pmatrix} 3 \\ 0 \\ -1 \\ 2 \end{pmatrix}, \mathbf{b}_3 = \begin{pmatrix} 2 \\ 3 \\ -2 \\ 1 \end{pmatrix}.$$

## 1.4 Pivoting and Permutations

The method of Gaussian Elimination presented so far applies only to regular matrices. But not every square matrix is regular; a simple class of examples is matrices whose upper left, i.e.,  $(1, 1)$ , entry is zero, and so cannot serve as the first pivot. More generally, the algorithm cannot proceed whenever a zero entry appears in the current pivot position on the diagonal. What then to do? The answer requires revisiting the source of the method.

Consider, as a specific example, the linear system

$$\begin{aligned} 2y + z &= 2, \\ 2x + 6y + z &= 7, \\ x + y + 4z &= 3. \end{aligned} \tag{1.29}$$

The augmented coefficient matrix is

$$\left( \begin{array}{ccc|c} 0 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array} \right).$$

In this case, the  $(1, 1)$  entry is 0, and so is not a legitimate pivot. The problem, of course, is that the first variable  $x$  does not appear in the first equation, and so we cannot use it to eliminate  $x$  in the other two equations. But this “problem” is actually a bonus — we already have an equation with only two variables in it, and so we need to eliminate  $x$  from only one of the other two equations. To be systematic, we rewrite the system in a different order,

$$2x + 6y + z = 7,$$

$$2y + z = 2,$$

$$x + y + 4z = 3,$$

by interchanging the first two equations. In other words, we employ

**Linear System Operation #2:** Interchange two equations.

Clearly, this operation does not change the solution and so produces an equivalent linear system. In our case, the augmented coefficient matrix,

$$\left( \begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 2 & 1 & 2 \\ 1 & 1 & 4 & 3 \end{array} \right),$$

can be obtained from the original by performing the second type of row operation:

**Elementary Row Operation #2:** Interchange two rows of the matrix.

The new nonzero upper left entry, 2, can now serve as the first pivot, and we may continue to apply elementary row operations of type #1 to reduce our matrix to upper triangular form. For this particular example, we eliminate the remaining nonzero entry in the first column by subtracting  $\frac{1}{2}$  the first row from the last:

$$\left( \begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 2 & 1 & 2 \\ 0 & -2 & \frac{7}{2} & -\frac{1}{2} \end{array} \right).$$

The (2, 2) entry serves as the next pivot. To eliminate the nonzero entry below it, we add the second to the third row:

$$\left( \begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 2 & 1 & 2 \\ 0 & 0 & \frac{9}{2} & \frac{3}{2} \end{array} \right).$$

We have now placed the system in upper triangular form, with the three pivots 2, 2, and  $\frac{9}{2}$  along the diagonal. Back Substitution produces the solution  $x = \frac{5}{6}$ ,  $y = \frac{5}{6}$ ,  $z = \frac{1}{3}$ .

The row interchange that is required when a zero shows up in the diagonal pivot position is known as *pivoting*. Later, in Section 1.7, we will discuss practical reasons for pivoting even when a diagonal entry is nonzero. Let us distinguish the class of matrices that can be reduced to upper triangular form by Gaussian Elimination with pivoting. These matrices will prove to be of fundamental importance throughout linear algebra.

**Definition 1.6.** A square matrix is called *nonsingular* if it can be reduced to upper triangular form with all non-zero elements on the diagonal — the pivots — by elementary row operations of types 1 and 2.

In contrast, a *singular* square matrix cannot be reduced to such upper triangular form by such row operations, because at some stage in the elimination procedure the diagonal entry and all the entries below it are zero. Every regular matrix is nonsingular, but, as we just saw, not every nonsingular matrix is regular. Uniqueness of solutions is the key defining characteristic of nonsingularity.

**Theorem 1.7.** A linear system  $A\mathbf{x} = \mathbf{b}$  has a unique solution for *every* choice of right-hand side  $\mathbf{b}$  if and only if its coefficient matrix  $A$  is square and nonsingular.

We are able to prove the “if” part of this theorem, since nonsingularity implies reduction to an equivalent upper triangular form that has the same solutions as the original system.

The unique solution to the system is then found by Back Substitution. The “only if” part will be proved in Section 1.8.

The revised version of the Gaussian Elimination algorithm, valid for all nonsingular coefficient matrices, is implemented by the accompanying pseudocode program. The starting point is the augmented matrix  $M = (A \mid \mathbf{b})$  representing the linear system  $A\mathbf{x} = \mathbf{b}$ . After successful termination of the program, the result is an augmented matrix in upper triangular form  $M = (U \mid \mathbf{c})$  representing the equivalent linear system  $U\mathbf{x} = \mathbf{c}$ . One then uses Back Substitution to determine the solution  $\mathbf{x}$  to the linear system.

---

*Gaussian Elimination — Nonsingular Case*

---

```

start
  for j = 1 to n
    if  $m_{kj} = 0$  for all  $k \geq j$ , stop; print “ $A$  is singular”
    if  $m_{jj} = 0$  but  $m_{kj} \neq 0$  for some  $k > j$ , switch rows  $k$  and  $j$ 
    for i =  $j + 1$  to n
      set  $l_{ij} = m_{ij}/m_{jj}$ 
      add  $-l_{ij}$  times row  $j$  to row  $i$  of  $M$ 
    next i
  next j
end

```

---

**Remark.** When performing the algorithm using exact arithmetic, when pivoting is required it does not matter which row  $k$  one chooses to switch with row  $j$ , as long as it lies below and the  $(k, j)$  entry is nonzero. When dealing with matters involving numerical precision and round off errors, there are some practical rules of thumb to be followed to maintain accuracy in the intervening computations. These will be discussed in Section 1.7.

---

**Exercises**

1.4.1. Determine whether the following matrices are singular or nonsingular:

$$(a) \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}, (b) \begin{pmatrix} -1 & 2 \\ 4 & -8 \end{pmatrix}, (c) \begin{pmatrix} 0 & 1 & 2 \\ -1 & 1 & 3 \\ 2 & -2 & 0 \end{pmatrix}, (d) \begin{pmatrix} 1 & 1 & 3 \\ 2 & 2 & 2 \\ 3 & -1 & 1 \end{pmatrix}, (e) \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix},$$

$$(f) \begin{pmatrix} -1 & 1 & 0 & -3 \\ 2 & -2 & 4 & 0 \\ 1 & -2 & 2 & -1 \\ 0 & 1 & 0 & 1 \end{pmatrix}, (g) \begin{pmatrix} 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -2 \\ 2 & 0 & 2 & 0 \end{pmatrix}, (h) \begin{pmatrix} 1 & -2 & 0 & 2 \\ 4 & 1 & -1 & -1 \\ -8 & -1 & 2 & 1 \\ -4 & -1 & 1 & 2 \end{pmatrix}.$$

1.4.2. Classify the following matrices as (i) regular, (ii) nonsingular, and/or (iii) singular:

$$(a) \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}, (b) \begin{pmatrix} 3 & -2 & 1 \\ -1 & 4 & 4 \\ 2 & 2 & 5 \end{pmatrix}, (c) \begin{pmatrix} 1 & -2 & 3 \\ -2 & 4 & -1 \\ 3 & -1 & 2 \end{pmatrix}, (d) \begin{pmatrix} 1 & 3 & -3 & 0 \\ -1 & 0 & -1 & 2 \\ 3 & -2 & 6 & 1 \\ 2 & -1 & 3 & 5 \end{pmatrix}.$$

1.4.3. Find the equation  $z = ax + by + c$  for the plane passing through the three points  $\mathbf{p}_1 = (0, 2, -1)$ ,  $\mathbf{p}_2 = (-2, 4, 3)$ ,  $\mathbf{p}_3 = (2, -1, -3)$ .

- 1.4.4. Show that a  $2 \times 2$  matrix  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is (a) nonsingular if and only if  $ad - bc \neq 0$ ,  
 (b) regular if and only if  $ad - bc \neq 0$  and  $a \neq 0$ .

- 1.4.5. Solve the following systems of equations by Gaussian Elimination:

$$\begin{array}{lll}
 x_1 - 2x_2 + 2x_3 = 15, & 2x_1 - x_2 = 1, & x_2 - x_3 = 4, \\
 \text{(a)} \quad x_1 - 2x_2 + x_3 = 10, & \text{(b)} \quad -4x_1 + 2x_2 - 3x_3 = -8, & \text{(c)} \quad -2x_1 - 5x_2 = 2, \\
 2x_1 - x_2 - 2x_3 = -10. & x_1 - 3x_2 + x_3 = 5. & x_1 + x_3 = -8. \\
 \text{(d)} \quad x - y + z - w = 0, \quad -2x + 2y - z + w = 2, & \text{(e)} \quad -3x_2 + 2x_3 = 0, \quad x_3 - x_4 = 2, \\
 -4x + 4y + 3z = 5, \quad x - 3y + w = 4. & x_1 - 2x_3 = -1, \quad -4x_1 + 7x_4 = -5. 
 \end{array}$$

- 1.4.6. *True or false:* A singular matrix cannot be regular.

- 1.4.7. *True or false:* A square matrix that has a column with all 0 entries is singular. What can you say about a linear system that has such a coefficient matrix?

- ◇ 1.4.8. Explain why the solution to the homogeneous system  $A \mathbf{x} = \mathbf{0}$  with nonsingular coefficient matrix is  $\mathbf{x} = \mathbf{0}$ .

- 1.4.9. Write out the details of the proof of the “if” part of Theorem 1.7: if  $A$  is nonsingular, then the linear system  $A \mathbf{x} = \mathbf{b}$  has a unique solution for every  $\mathbf{b}$ .

## Permutations and Permutation Matrices

As with the first type of elementary row operation, row interchanges can be accomplished by multiplication by a second type of elementary matrix, which is found by applying the row operation to the identity matrix of the appropriate size. For instance, interchanging rows 1 and 2 of the  $3 \times 3$  identity matrix produces the elementary interchange matrix

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad \text{The result } PA \text{ of multiplying any 3-rowed matrix } A \text{ on the left by } P \text{ is}$$

the same as interchanging the first two rows of  $A$ . For instance,

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{pmatrix}.$$

Multiple row interchanges are accomplished by combining such elementary interchange matrices. Each such combination of row interchanges uniquely corresponds to what is called a permutation matrix.

**Definition 1.8.** A *permutation matrix* is a matrix obtained from the identity matrix by any combination of row interchanges.

In particular, applying a row interchange to a permutation matrix produces another permutation matrix. The following result is easily established.

**Lemma 1.9.** A matrix  $P$  is a permutation matrix if and only if each row of  $P$  contains all 0 entries except for a single 1, and, in addition, each column of  $P$  also contains all 0 entries except for a single 1.

In general, if, in the permutation matrix  $P$ , a 1 appears in position  $(i, j)$ , then multiplication by  $P$  will move the  $j^{\text{th}}$  row of  $A$  into the  $i^{\text{th}}$  row of the product  $PA$ .

**Example 1.10.** There are six different  $3 \times 3$  permutation matrices, namely

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \quad (1.30)$$

These have the following effects: if  $A$  is a matrix with row vectors  $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ , then multiplication on the left by each of the six permutation matrices produces, respectively,

$$\begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_2 \\ \mathbf{r}_3 \\ \mathbf{r}_1 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_3 \\ \mathbf{r}_1 \\ \mathbf{r}_2 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_2 \\ \mathbf{r}_1 \\ \mathbf{r}_3 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_3 \\ \mathbf{r}_2 \\ \mathbf{r}_1 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_3 \\ \mathbf{r}_2 \end{pmatrix}. \quad (1.31)$$

Thus, the first permutation matrix, which is the identity, does nothing — the *identity permutation*. The fourth, fifth, sixth represent row interchanges. The second and third are non-elementary permutations, and can be realized by a pair of successive row interchanges.

In general, any rearrangement of a finite ordered collection of objects is called a *permutation*. Thus, the 6 permutation matrices (1.30) produce the 6 possible permutations (1.31) of the rows of a  $3 \times 3$  matrix. In general, if a permutation  $\pi$  rearranges the integers  $(1, \dots, n)$  to form  $(\pi(1), \dots, \pi(n))$ , then the corresponding permutation matrix  $P = P_\pi$  that maps row  $\mathbf{r}_i$  to row  $\mathbf{r}_{\pi(i)}$  will have 1's in positions  $(i, \pi(i))$  for  $i = 1, \dots, n$  and zeros everywhere else. For example, the second permutation matrix in (1.30) corresponds to the permutation with  $\pi(1) = 2, \pi(2) = 3, \pi(3) = 1$ . Keep in mind that  $\pi(1), \dots, \pi(n)$  is merely a rearrangement of the integers  $1, \dots, n$ , so that  $1 \leq \pi(i) \leq n$  and  $\pi(i) \neq \pi(j)$  when  $i \neq j$ .

An elementary combinatorial argument proves that there is a total of

$$n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1 \quad (1.32)$$

different permutations of  $(1, \dots, n)$ , and hence the same number of permutation matrices of size  $n \times n$ . Moreover, the product  $P = P_1 P_2$  of any two permutation matrices is also a permutation matrix, and corresponds to the composition of the two permutations, meaning one permutes according to  $P_2$  and then permutes the result according to  $P_1$ . An important point is that multiplication of permutation matrices is *noncommutative* — the order in which one permutes makes a difference. Switching the first and second rows, and then switching the second and third rows, *does not* have the same effect as first switching the second and third rows and then switching the first and second rows!

## Exercises

1.4.10. Write down the elementary  $4 \times 4$  permutation matrix (a)  $P_1$  that permutes the second and fourth rows, and (b)  $P_2$  that permutes the first and fourth rows. (c) Do  $P_1$  and  $P_2$  commute? (d) Explain what the matrix products  $P_1 P_2$  and  $P_2 P_1$  do to a  $4 \times 4$  matrix.

1.4.11. Write down the permutation matrix  $P$  such that

$$(a) P \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} v \\ w \\ u \end{pmatrix}, \quad (b) P \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} d \\ c \\ a \\ b \end{pmatrix}, \quad (c) P \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} b \\ a \\ d \\ c \end{pmatrix}, \quad (d) P \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} x_4 \\ x_5 \\ x_1 \\ x_3 \\ x_2 \end{pmatrix}.$$

1.4.12. Construct a multiplication table that shows all possible products of the  $3 \times 3$  permutation matrices (1.30). List all pairs that commute.

1.4.13. Write down all  $4 \times 4$  permutation matrices that (a) fix the third row of a  $4 \times 4$  matrix  $A$ ; (b) take the third row to the fourth row; (c) interchange the second and third rows.

1.4.14. *True or false:* (a) Every elementary permutation matrix satisfies  $P^2 = I$ . (b) Every permutation matrix satisfies  $P^2 = I$ . (c) A matrix that satisfies  $P^2 = I$  is necessarily a permutation matrix.

1.4.15. (a) Let  $P$  and  $Q$  be  $n \times n$  permutation matrices and  $\mathbf{v} \in \mathbb{R}^n$  a vector. Under what conditions does the equation  $P\mathbf{v} = Q\mathbf{v}$  imply that  $P = Q$ ? (b) Answer the same question when  $PA = QA$ , where  $A$  is an  $n \times k$  matrix.

1.4.16. Let  $P$  be the  $3 \times 3$  permutation matrix such that the product  $PA$  permutes the first and third rows of the  $3 \times 3$  matrix  $A$ . (a) Write down  $P$ . (b) *True or false:* The product  $AP$  is obtained by permuting the first and third columns of  $A$ .

(c) Does the same conclusion hold for every permutation matrix: is the effect of  $PA$  on the rows of a square matrix  $A$  the same as the effect of  $AP$  on the columns of  $A$ ?

◇ 1.4.17. A common notation for a permutation  $\pi$  of the integers  $\{1, \dots, m\}$  is as a  $2 \times m$  matrix  $\begin{pmatrix} 1 & 2 & 3 & \dots & m \\ \pi(1) & \pi(2) & \pi(3) & \dots & \pi(m) \end{pmatrix}$ , indicating that  $\pi$  takes  $i$  to  $\pi(i)$ . (a) Show that such a permutation corresponds to the permutation matrix with 1's in positions  $(\pi(j), j)$  for  $j = 1, \dots, m$ . (b) Write down the permutation matrices corresponding to the following permutations: (i)  $\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 3 & 1 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}$ , (iv)  $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 3 & 2 & 1 \end{pmatrix}$ . Which are elementary matrices? (c) Write down, using the preceding notation, the permutations corresponding to the following permutation matrices:

$$(i) \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad (ii) \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad (iii) \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad (iv) \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

◇ 1.4.18. Justify the statement that there are  $n!$  different  $n \times n$  permutation matrices.

1.4.19. Consider the following combination of elementary row operations of type #1: (i) Add row  $i$  to row  $j$ . (ii) Subtract row  $j$  from row  $i$ . (iii) Add row  $i$  to row  $j$  again. Prove that the net effect is to interchange  $-1$  times row  $i$  with row  $j$ . Thus, we can *almost* produce an elementary row operation of type #2 by a combination of elementary row operations of type #1. Lest you be tempted to try, Exercise 1.9.16 proves that one *cannot* produce a bona fide row interchange by a combination of elementary row operations of type #1.

1.4.20. What is the effect of permuting the *columns* of its coefficient matrix on a linear system?

## The Permuted LU Factorization

As we now know, every nonsingular matrix  $A$  can be reduced to upper triangular form by elementary row operations of types #1 and #2. The row interchanges merely reorder the equations. If one performs all of the required row interchanges in advance, then the elimination algorithm can proceed without requiring any further pivoting. Thus, the matrix obtained by permuting the rows of  $A$  in the prescribed manner is regular. In other words, if  $A$  is a nonsingular matrix, then there is a permutation matrix  $P$  such that the product  $PA$  is regular, and hence admits an *LU* factorization. As a result, we deduce the general *permuted LU factorization*

$$PA = LU, \tag{1.33}$$

where  $P$  is a permutation matrix,  $L$  is lower unitriangular, and  $U$  is upper triangular with the pivots on the diagonal. For instance, in the preceding example, we permuted the first and second rows, and hence equation (1.33) has the explicit form

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & \frac{9}{2} \end{pmatrix}.$$

We have now established the following generalization of Theorem 1.3.

**Theorem 1.11.** Let  $A$  be an  $n \times n$  matrix. Then the following conditions are equivalent:

- (i)  $A$  is nonsingular.
- (ii)  $A$  has  $n$  nonzero pivots.
- (iii)  $A$  admits a permuted  $LU$  factorization:  $PA = LU$ .

A practical method to construct a permuted  $LU$  factorization of a given matrix  $A$  would proceed as follows. First set up  $P = L = I$  as  $n \times n$  identity matrices. The matrix  $P$  will keep track of the permutations performed during the Gaussian Elimination process, while the entries of  $L$  below the diagonal are gradually replaced by the negatives of the multiples used in the corresponding row operations of type #1. Each time two rows of  $A$  are interchanged, the same two rows of  $P$  will be interchanged. Moreover, any pair of entries that both lie *below* the diagonal in these same two rows of  $L$  must also be interchanged, while entries lying on and above its diagonal need to stay in their place. At a successful conclusion to the procedure,  $A$  will have been converted into the upper triangular matrix  $U$ , while  $L$  and  $P$  will assume their final form. Here is an illustrative example.

**Example 1.12.** Our goal is to produce a permuted  $LU$  factorization of the matrix

$$A = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 2 & 4 & -2 & -1 \\ -3 & -5 & 6 & 1 \\ -1 & 2 & 8 & -2 \end{pmatrix}.$$

To begin the procedure, we apply row operations of type #1 to eliminate the entries below the first pivot. The updated matrices<sup>†</sup> are

$$A = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 3 & 1 \\ 0 & 4 & 7 & -2 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where  $L$  keeps track of the row operations, and we initialize  $P$  to be the identity matrix. The  $(2,2)$  entry of the new  $A$  is zero, and so we interchange its second and third rows, leading to

$$A = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 4 & 7 & -2 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

---

<sup>†</sup> Here, we are adopting computer programming conventions, where updates of a matrix are all given the same name.

We interchanged the same two rows of  $P$ , while in  $L$  we only interchanged the already computed entries in its second and third rows that lie in its first column below the diagonal. We then eliminate the nonzero entry lying below the  $(2, 2)$  pivot, leading to

$$A = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -5 & -6 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -1 & 4 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

A final row interchange places the matrix in upper triangular form:

$$U = A = \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & -5 & -6 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ -1 & 4 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Again, we performed the same row interchange on  $P$ , while interchanging only the third and fourth row entries of  $L$  that lie below the diagonal. You can verify that

$$PA = \begin{pmatrix} 1 & 2 & -1 & 0 \\ -3 & -5 & 6 & 1 \\ -1 & 2 & 8 & -2 \\ 2 & 4 & -2 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ -1 & 4 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & -5 & -6 \\ 0 & 0 & 0 & -1 \end{pmatrix} = LU, \quad (1.34)$$

as promised. Thus, by rearranging the equations in the order first, third, fourth, second, as prescribed by  $P$ , we obtain an equivalent linear system whose coefficient matrix  $PA$  is regular, in accordance with Theorem 1.11.

Once the permuted  $LU$  factorization is established, the solution to the original system  $A\mathbf{x} = \mathbf{b}$  is obtained by applying the same Forward and Back Substitution algorithm presented above. Explicitly, we first multiply the system  $A\mathbf{x} = \mathbf{b}$  by the permutation matrix, leading to

$$PA\mathbf{x} = P\mathbf{b} = \hat{\mathbf{b}}, \quad (1.35)$$

whose right-hand side  $\hat{\mathbf{b}}$  has been obtained by permuting the entries of  $\mathbf{b}$  in the same fashion as the rows of  $A$ . We then solve the two triangular systems

$$L\mathbf{c} = \hat{\mathbf{b}} \quad \text{and} \quad U\mathbf{x} = \mathbf{c} \quad (1.36)$$

by, respectively, Forward and Back Substitution, as before.

**Example 1.12 (continued).** Suppose we wish to solve the linear system

$$\begin{pmatrix} 1 & 2 & -1 & 0 \\ 2 & 4 & -2 & -1 \\ -3 & -5 & 6 & 1 \\ -1 & 2 & 8 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 3 \\ 0 \end{pmatrix}.$$

In view of the  $PA = LU$  factorization established in (1.34), we need only solve the two auxiliary lower and upper triangular systems (1.36). The lower triangular system is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 \\ -1 & 4 & 1 & 0 \\ 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 0 \\ -1 \end{pmatrix};$$

whose right-hand side was obtained by applying the permutation matrix  $P$  to the right-hand side of the original system. Its solution, namely  $a = 1$ ,  $b = 6$ ,  $c = -23$ ,  $d = -3$ , is obtained through Forward Substitution. The resulting upper triangular system is

$$\begin{pmatrix} 1 & 2 & -1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & -5 & -6 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 6 \\ -23 \\ -3 \end{pmatrix}.$$

Its solution,  $w = 3$ ,  $z = 1$ ,  $y = 0$ ,  $x = 2$ , which is also the solution to the original system, is easily obtained by Back Substitution.

## Exercises

1.4.21. For each of the listed matrices  $A$  and vectors  $\mathbf{b}$ , find a permuted  $LU$  factorization of the matrix, and use your factorization to solve the system  $A\mathbf{x} = \mathbf{b}$ . (a)  $\begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}$ ,  $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ ,

$$(b) \begin{pmatrix} 0 & 0 & -4 \\ 1 & 2 & 3 \\ 0 & 1 & 7 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, (c) \begin{pmatrix} 0 & 1 & -3 \\ 0 & 2 & 3 \\ 1 & 0 & 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, (d) \begin{pmatrix} 1 & 2 & -1 & 0 \\ 3 & 6 & 2 & -1 \\ 1 & 1 & -7 & 2 \\ 1 & -1 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 3 \end{pmatrix},$$

$$(e) \begin{pmatrix} 0 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 \\ 1 & 4 & -1 & 2 \\ 7 & -1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} -1 \\ -4 \\ 0 \\ 5 \end{pmatrix}, (f) \begin{pmatrix} 0 & 0 & 2 & 3 & 4 \\ 0 & 1 & -7 & 2 & 3 \\ 1 & 4 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 7 & 3 \end{pmatrix}, \begin{pmatrix} -3 \\ -2 \\ 0 \\ 0 \\ -7 \end{pmatrix}.$$

1.4.22. For each of the following linear systems find a permuted  $LU$  factorization of the coefficient matrix and then use it to solve the system by Forward and Back Substitution.

$$\begin{array}{l} 4x_1 - 4x_2 + 2x_3 = 1, \\ (a) \begin{array}{l} -3x_1 + 3x_2 + x_3 = 3, \\ -3x_1 + x_2 - 2x_3 = -5. \end{array} \end{array} \quad \begin{array}{l} y - z + w = 0, \\ (b) \begin{array}{l} x - y + z - 3w = 2, \\ x + 2y - z + w = 4. \end{array} \end{array} \quad \begin{array}{l} x - y + 2z + w = 0, \\ (c) \begin{array}{l} -x + y - 3z = 1, \\ x - y + 4z - 3w = 2, \\ x + 2y - z + w = 4. \end{array} \end{array}$$

◇ 1.4.23. (a) Explain why

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 3 \\ 2 & -1 & 1 \\ 2 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 3 \\ 2 & -1 & 1 \\ 2 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix},$$

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 3 \\ 2 & -1 & 1 \\ 2 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & -2 \end{pmatrix},$$

are all legitimate permuted  $LU$  factorizations of the same matrix. List the elementary row operations that are being used in each case.

$$(b) \text{ Use each of the factorizations to solve the linear system } \begin{pmatrix} 0 & 1 & 3 \\ 2 & -1 & 1 \\ 2 & -2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -5 \\ -1 \\ 0 \end{pmatrix}.$$

Do you always obtain the same result? Explain why or why not.

1.4.24. (a) Find three different permuted  $LU$  factorizations of the matrix  $A = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & -1 \\ 1 & 1 & 3 \end{pmatrix}$ .  
 (b) How many different permuted  $LU$  factorizations does  $A$  have?

1.4.25. What is the maximal number of permuted  $LU$  factorizations a regular  $3 \times 3$  matrix can have? Give an example of such a matrix.

1.4.26. *True or false:* The pivots of a nonsingular matrix are uniquely defined.

♣ 1.4.27. (a) Write a pseudocode program implementing the algorithm for finding the permuted  $LU$  factorization of a matrix. (b) Program your algorithm and test it on the examples in Exercise 1.4.21.

## 1.5 Matrix Inverses

The inverse of a matrix is analogous to the reciprocal  $a^{-1} = 1/a$  of a nonzero scalar  $a \neq 0$ . We already encountered the inverses of matrices corresponding to elementary row operations. In this section, we will study inverses of general square matrices. We begin with the formal definition.

**Definition 1.13.** Let  $A$  be a square matrix of size  $n \times n$ . An  $n \times n$  matrix  $X$  is called the *inverse* of  $A$  if it satisfies

$$XA = I = AX, \quad (1.37)$$

where  $I = I_n$  is the  $n \times n$  identity matrix. The inverse of  $A$  is commonly denoted by  $A^{-1}$ .

**Remark.** Noncommutativity of matrix multiplication requires that we impose both conditions in (1.37) in order to properly define an inverse to the matrix  $A$ . The first condition,  $XA = I$ , says that  $X$  is a *left inverse*, while the second,  $AX = I$ , requires that  $X$  also be a *right inverse*. Rectangular matrices might have either a left inverse or a right inverse, but, as we shall see, *only* square matrices have both, and so only square matrices can have full-fledged inverses. However, not every square matrix has an inverse. Indeed, not every scalar has an inverse:  $0^{-1} = 1/0$  is not defined, since the equation  $0x = 1$  has no solution.

**Example 1.14.** Since

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

we conclude that when  $A = \begin{pmatrix} 1 & 2 & -1 \\ -3 & 1 & 2 \\ -2 & 2 & 1 \end{pmatrix}$ , then  $A^{-1} = \begin{pmatrix} 3 & 4 & -5 \\ 1 & 1 & -1 \\ 4 & 6 & -7 \end{pmatrix}$ . Observe that there is no obvious way to anticipate the entries of  $A^{-1}$  from the entries of  $A$ .

**Example 1.15.** Let us compute the inverse  $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$ , when it exists, of a general  $2 \times 2$  matrix  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . The right inverse condition

$$AX = \begin{pmatrix} ax + bz & ay + bw \\ cx + dz & cy + dw \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$$

holds if and only if  $x, y, z, w$  satisfy the linear system

$$\begin{aligned} ax + bz &= 1, & ay + bw &= 0, \\ cx + dz &= 0, & cy + dw &= 1. \end{aligned}$$

Solving by Gaussian Elimination (or directly), we find

$$x = \frac{d}{ad - bc}, \quad y = -\frac{b}{ad - bc}, \quad z = -\frac{c}{ad - bc}, \quad w = \frac{a}{ad - bc},$$

provided the common denominator  $ad - bc \neq 0$  does not vanish. Therefore, the matrix

$$X = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

forms a right inverse to  $A$ . However, a short computation shows that it also defines a left inverse:

$$XA = \begin{pmatrix} xa + yc & xb + yd \\ za + wc & zb + wd \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I,$$

and hence  $X = A^{-1}$  is the inverse of  $A$ .

The denominator appearing in the preceding formulas has a special name; it is called the *determinant* of the  $2 \times 2$  matrix  $A$ , and denoted by

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc. \quad (1.38)$$

Thus, the determinant of a  $2 \times 2$  matrix is the product of the diagonal entries minus the product of the off-diagonal entries. (Determinants of larger square matrices will be discussed in Section 1.9.) Thus, the  $2 \times 2$  matrix  $A$  is invertible, with

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \quad (1.39)$$

if and only if  $\det A \neq 0$ . For example, if  $A = \begin{pmatrix} 1 & 3 \\ -2 & -4 \end{pmatrix}$ , then  $\det A = 2 \neq 0$ . We conclude that  $A$  has an inverse, which, by (1.39), is  $A^{-1} = \frac{1}{2} \begin{pmatrix} -4 & -3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & -\frac{3}{2} \\ 1 & \frac{1}{2} \end{pmatrix}$ .

**Example 1.16.** We already learned how to find the inverse of an elementary matrix of type #1: we just negate the one nonzero off-diagonal entry. For example, if

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \quad \text{then} \quad E^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

This is because the inverse of the elementary row operation that adds twice the first row to the third row is the operation of subtracting twice the first row from the third row.

**Example 1.17.** Let  $P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$  denote the elementary matrix that has the effect

of interchanging rows 1 and 2 of a 3-rowed matrix. Then  $P^2 = I$ , since performing the interchange twice returns us to where we began. This implies that  $P^{-1} = P$  is its own inverse. Indeed, the same result holds for all elementary permutation matrices that correspond to row operations of type #2. However, it is not true for more general permutation matrices.

The following fundamental result will be established later in this chapter.

**Theorem 1.18.** A square matrix has an inverse if and only if it is nonsingular.

Consequently, an  $n \times n$  matrix will have an inverse if and only if it can be reduced to upper triangular form, with  $n$  nonzero pivots on the diagonal, by a combination of elementary row operations. Indeed, “invertible” is often used as a synonym for “nonsingular”. All other matrices are singular and do not have an inverse as defined above. Before attempting to prove Theorem 1.18, we need first to become familiar with some elementary properties of matrix inverses.

**Lemma 1.19.** The inverse of a square matrix, if it exists, is unique.

*Proof:* Suppose both  $X$  and  $Y$  satisfy (1.37), so

$$XA = I = AX \quad \text{and} \quad YA = I = AY.$$

Then, by associativity,

$$X = X I = X(A Y) = (X A) Y = I Y = Y. \quad \text{Q.E.D.}$$

Inverting a matrix twice brings us back to where we started.

**Lemma 1.20.** If  $A$  is an invertible matrix, then  $A^{-1}$  is also invertible and  $(A^{-1})^{-1} = A$ .

*Proof:* The matrix inverse equations  $A^{-1} A = I = A A^{-1}$  are sufficient to prove that  $A$  is the inverse of  $A^{-1}$ . Q.E.D.

**Lemma 1.21.** If  $A$  and  $B$  are invertible matrices of the same size, then their product,  $AB$ , is invertible, and

$$(AB)^{-1} = B^{-1}A^{-1}. \quad (1.40)$$

Note that the order of the factors is reversed under inversion.

*Proof:* Let  $X = B^{-1}A^{-1}$ . Then, by associativity,

$$\begin{aligned} X(AB) &= B^{-1}A^{-1}AB = B^{-1}IB = B^{-1}B = I, \\ (AB)X &= ABB^{-1}A^{-1} = AIA^{-1} = AA^{-1} = I. \end{aligned}$$

Thus  $X$  is both a left and a right inverse for the product matrix  $AB$ .

Q.E.D.

**Example 1.22.** One verifies, directly, that the inverse of  $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$  is  $A^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ , while the inverse of  $B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  is  $B^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ . Therefore, the inverse of their product  $C = AB = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ -1 & 0 \end{pmatrix}$  is given by  $C^{-1} = B^{-1}A^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & -2 \end{pmatrix}$ .

We can straightforwardly generalize the preceding result. The inverse of a  $k$ -fold product of invertible matrices is the product of their inverses, *in the reverse order*:

$$(A_1 A_2 \cdots A_{k-1} A_k)^{-1} = A_k^{-1} A_{k-1}^{-1} \cdots A_2^{-1} A_1^{-1}. \quad (1.41)$$

**Warning.** In general,  $(A + B)^{-1} \neq A^{-1} + B^{-1}$ . Indeed, this equation is not even true for scalars ( $1 \times 1$  matrices)!

## Exercises

1.5.1. Verify by direct multiplication that the following matrices are inverses, i.e., both

conditions in (1.37) hold: (a)  $A = \begin{pmatrix} 2 & -3 \\ -1 & -1 \end{pmatrix}$ ,  $A^{-1} = \begin{pmatrix} -1 & -3 \\ 1 & 2 \end{pmatrix}$ ; (b)  $A = \begin{pmatrix} 2 & 1 & 1 \\ 3 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}$ ,

$A^{-1} = \begin{pmatrix} 3 & -1 & -1 \\ -4 & 2 & 1 \\ -1 & 0 & 1 \end{pmatrix}$ ; (c)  $A = \begin{pmatrix} -1 & 3 & 2 \\ 2 & 2 & -1 \\ -2 & 1 & 3 \end{pmatrix}$ ,  $A^{-1} = \begin{pmatrix} -1 & 1 & 1 \\ \frac{4}{7} & -\frac{1}{7} & -\frac{3}{7} \\ -\frac{6}{7} & \frac{5}{7} & \frac{8}{7} \end{pmatrix}$ .

1.5.2. Let  $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 3 \\ 1 & -1 & -8 \end{pmatrix}$ . Find the right inverse of  $A$  by setting up and solving the linear system  $AX = I$ . Verify that the resulting matrix  $X$  is also a left inverse.

1.5.3. Write down the inverse of each of the following elementary matrices: (a)  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ ,

(b)  $\begin{pmatrix} 1 & 0 \\ 5 & 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}$ , (e)  $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 6 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$ .

1.5.4. Show that the inverse of  $L = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & 0 & 1 \end{pmatrix}$  is  $L^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ -b & 0 & 1 \end{pmatrix}$ . However, the inverse of  $M = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}$  is not  $\begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ -b & -c & 1 \end{pmatrix}$ . What is  $M^{-1}$ ?

1.5.5. Explain why a matrix with a row of all zeros does not have an inverse.

1.5.6. (a) Write down the inverse of the matrices  $A = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$  and  $B = \begin{pmatrix} 1 & -1 \\ 1 & 2 \end{pmatrix}$ . (b) Write down the product matrix  $C = AB$  and its inverse  $C^{-1}$  using the inverse product formula.

1.5.7. (a) Find the inverse of the *rotation matrix*  $R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$ , where  $\theta \in \mathbb{R}$ .

(b) Use your result to solve the system  $x = a \cos \theta - b \sin \theta$ ,  $y = a \sin \theta + b \cos \theta$ , for  $a$  and  $b$  in terms of  $x$  and  $y$ . (c) Prove that, for all  $a \in \mathbb{R}$  and  $0 < \theta < \pi$ , the matrix  $R_\theta - aI$  has an inverse.

1.5.8. (a) Write down the inverses of each of the  $3 \times 3$  permutation matrices (1.30). (b) Which ones are their own inverses,  $P^{-1} = P$ ? (c) Can you find a non-elementary permutation matrix  $P$  that is its own inverse:  $P^{-1} = P$ ?

1.5.9. Find the inverse of the following permutation matrices:

(a)  $\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$ , (b)  $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ .

1.5.10. Explain how to write down the inverse permutation using the notation of Exercise 1.4.17. Apply your method to the examples in Exercise 1.5.9, and check the result by verifying that it produces the inverse permutation matrix.

1.5.11. Find all real  $2 \times 2$  matrices that are their own inverses:  $A^{-1} = A$ .

1.5.12. Show that if a square matrix  $A$  satisfies  $A^2 - 3A + I = 0$ , then  $A^{-1} = 3I - A$ .

1.5.13. Prove that if  $c \neq 0$  is any nonzero scalar and  $A$  is an invertible matrix, then the scalar product matrix  $cA$  is invertible, and  $(cA)^{-1} = \frac{1}{c}A^{-1}$ .

1.5.14. Show that  $A = \begin{pmatrix} 0 & a & 0 & 0 & 0 \\ b & 0 & c & 0 & 0 \\ 0 & d & 0 & e & 0 \\ 0 & 0 & f & 0 & g \\ 0 & 0 & 0 & h & 0 \end{pmatrix}$  is not invertible for any value of the entries.

1.5.15. Show that if  $A$  is a nonsingular matrix, so is every power  $A^n$ .

1.5.16. Prove that a diagonal matrix  $D = \text{diag}(d_1, \dots, d_n)$  is invertible if and only if all its diagonal entries are nonzero, in which case  $D^{-1} = \text{diag}(1/d_1, \dots, 1/d_n)$ .

1.5.17. Prove that if  $U$  is a nonsingular upper triangular matrix, then the diagonal entries of  $U^{-1}$  are the reciprocals of the diagonal entries of  $U$ .

◇ 1.5.18.(a) Let  $U$  be a  $m \times n$  matrix and  $V$  an  $n \times m$  matrix, such that the  $m \times m$  matrix  $I_m + UV$  is invertible. Prove that  $I_n + VU$  is also invertible, and is given by

$$(I_n + VU)^{-1} = I_n - V(I_m + UV)^{-1}U.$$

(b) The *Sherman–Morrison–Woodbury formula* generalizes this identity to

$$(A + VBU)^{-1} = A^{-1} - A^{-1}V(B^{-1} + UA^{-1}V)^{-1}UA^{-1}. \quad (1.42)$$

Explain what assumptions must be made on the matrices  $A, B, U, V$  for (1.42) to be valid.

◇ 1.5.19. Two matrices  $A$  and  $B$  are said to be *similar*, written  $A \sim B$ , if there exists an invertible matrix  $S$  such that  $B = S^{-1}AS$ . Prove: (a)  $A \sim A$ . (b) If  $A \sim B$ , then  $B \sim A$ . (c) If  $A \sim B$  and  $B \sim C$ , then  $A \sim C$ .

◇ 1.5.20.(a) A block matrix  $D = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$  is called *block diagonal* if  $A$  and  $B$  are square matrices, not necessarily of the same size, while the 0's are zero matrices of the appropriate sizes. Prove that  $D$  has an inverse if and only if both  $A$  and  $B$  do, and

$D^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix}$ . (b) Find the inverse of  $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$  and  $\begin{pmatrix} 1 & -1 & 0 & 0 \\ 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & 5 \end{pmatrix}$  by using this method.

1.5.21.(a) Show that  $B = \begin{pmatrix} 1 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix}$  is a left inverse of  $A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$ . (b) Show that

$A$  does not have a right inverse. (c) Can you find any other left inverses of  $A$ ?

1.5.22. Prove that the rectangular matrix  $A = \begin{pmatrix} 1 & 2 & -1 \\ 1 & 2 & 0 \end{pmatrix}$  has a right inverse, but no left inverse.

1.5.23.(a) Are there any nonzero real scalars that satisfy  $(a+b)^{-1} = a^{-1} + b^{-1}$ ?

(b) Are there any nonsingular real  $2 \times 2$  matrices that satisfy  $(A+B)^{-1} = A^{-1} + B^{-1}$ ?

## Gauss–Jordan Elimination

The principal algorithm used to compute the inverse of a nonsingular matrix is known as *Gauss–Jordan Elimination*, in honor of Gauss and Wilhelm Jordan, a nineteenth-century German engineer. A key fact is that, given that  $A$  is square, we need to solve only the right inverse equation

$$AX = I \quad (1.43)$$

in order to compute  $X = A^{-1}$ . The left inverse equation in (1.37), namely  $XA = I$ , will then follow as an automatic consequence. In other words, for square matrices, a right inverse is automatically a left inverse, and conversely! A proof will appear below.

The reader may well ask, then, why use both left and right inverse conditions in the original definition? There are several good reasons. First of all, a non-square matrix may satisfy one of the two conditions — having either a left inverse or a right inverse — but can never satisfy both. Moreover, even when we restrict our attention to square matrices, starting with only one of the conditions makes the logical development of the subject considerably more difficult, and not really worth the extra effort. Once we have established the basic properties of the inverse of a square matrix, we can then safely discard the superfluous left inverse condition. Finally, when we generalize the notion of an inverse to linear operators in Chapter 7, then, in contrast to the case of square matrices, we *cannot* dispense with either of the conditions.

Let us write out the individual columns of the right inverse equation (1.43). The  $j^{\text{th}}$  column of the  $n \times n$  identity matrix  $I$  is the vector  $\mathbf{e}_j$  that has a 1 in the  $j^{\text{th}}$  slot and 0's elsewhere, so

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \dots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \quad (1.44)$$

According to (1.11), the  $j^{\text{th}}$  column of the matrix product  $AX$  is equal to  $A\mathbf{x}_j$ , where  $\mathbf{x}_j$  denotes the  $j^{\text{th}}$  column of the inverse matrix  $X$ . Therefore, the single matrix equation (1.43) is equivalent to  $n$  linear systems

$$A\mathbf{x}_1 = \mathbf{e}_1, \quad A\mathbf{x}_2 = \mathbf{e}_2, \quad \dots \quad A\mathbf{x}_n = \mathbf{e}_n, \quad (1.45)$$

all having the same coefficient matrix. As such, to solve them we should form the  $n$  augmented matrices  $M_1 = (A \mid \mathbf{e}_1), \dots, M_n = (A \mid \mathbf{e}_n)$ , and then apply our Gaussian Elimination algorithm to each. But this would be a waste of effort. Since the coefficient matrix is the same, we will end up performing *identical* row operations on each augmented matrix. Clearly, it will be more efficient to combine them into one large augmented matrix  $M = (A \mid \mathbf{e}_1 \dots \mathbf{e}_n) = (A \mid I)$ , of size  $n \times (2n)$ , in which the right-hand sides  $\mathbf{e}_1, \dots, \mathbf{e}_n$  of our systems are placed into  $n$  different columns, which we then recognize as reassembling the columns of an  $n \times n$  identity matrix. We may then simultaneously apply our elementary row operations to reduce, if possible, the large augmented matrix so that its first  $n$  columns are in upper triangular form.

**Example 1.23.** For example, to find the inverse of the matrix  $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$ , we form the large augmented matrix

$$\left( \begin{array}{ccc|ccc} 0 & 2 & 1 & 1 & 0 & 0 \\ 2 & 6 & 1 & 0 & 1 & 0 \\ 1 & 1 & 4 & 0 & 0 & 1 \end{array} \right).$$

Applying the same sequence of elementary row operations as in Section 1.4, we first interchange the rows

$$\left( \begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 4 & 0 & 0 & 1 \end{array} \right),$$

and then eliminate the nonzero entries below the first pivot,

$$\left( \begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & -2 & \frac{7}{2} & 0 & -\frac{1}{2} & 1 \end{array} \right).$$

Next we eliminate the entry below the second pivot:

$$\left( \begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{9}{2} & 1 & -\frac{1}{2} & 1 \end{array} \right).$$

At this stage, we have reduced our augmented matrix to the form  $(U | C)$ , where  $U$  is upper triangular. This is equivalent to reducing the original  $n$  linear systems  $A\mathbf{x}_i = \mathbf{e}_i$  to  $n$  upper triangular systems  $U\mathbf{x}_i = \mathbf{c}_i$ . We can therefore perform  $n$  back substitutions to produce the solutions  $\mathbf{x}_i$ , which would form the individual columns of the inverse matrix  $X = (\mathbf{x}_1 \dots \mathbf{x}_n)$ . In the more common version of the Gauss–Jordan scheme, one instead continues to employ elementary row operations to fully reduce the augmented matrix. The goal is to produce an augmented matrix  $(I | X)$  in which the left-hand  $n \times n$  matrix has become the identity, while the right-hand matrix is the desired solution  $X = A^{-1}$ . Indeed,  $(I | X)$  represents the  $n$  trivial linear systems  $I\mathbf{x} = \mathbf{x}_i$  whose solutions  $\mathbf{x} = \mathbf{x}_i$  are the columns of the inverse matrix  $X$ .

Now, the identity matrix has 0's below the diagonal, just like  $U$ . It also has 1's along the diagonal, whereas  $U$  has the pivots (which are all nonzero) along the diagonal. Thus, the next phase in the reduction process is to make all the diagonal entries of  $U$  equal to 1. To proceed, we need to introduce the last, and least, of our linear systems operations.

**Linear System Operation #3:** Multiply an equation by a nonzero constant.

This operation clearly does not affect the solution, and so yields an equivalent linear system. The corresponding elementary row operation is:

**Elementary Row Operation #3:** Multiply a row of the matrix by a nonzero scalar.

Dividing the rows of the upper triangular augmented matrix  $(U | C)$  by the diagonal pivots of  $U$  will produce a matrix of the form  $(V | B)$ , where  $V$  is *upper unitriangular*, meaning it has all 1's along the diagonal. In our particular example, the result of these three elementary row operations of type #3 is

$$\left( \begin{array}{ccc|ccc} 1 & 3 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array} \right),$$

where we multiplied the first and second rows by  $\frac{1}{2}$  and the third row by  $\frac{2}{9}$ .

We are now over halfway towards our goal. We need only make the entries above the diagonal of the left-hand matrix equal to zero. This can be done by elementary row operations of type #1, but now we work backwards. First, we eliminate the nonzero entries in the third column lying above the  $(3, 3)$  entry by subtracting one half the third row from the second and also from the first:

$$\left( \begin{array}{ccc|ccc} 1 & 3 & 0 & -\frac{1}{9} & \frac{5}{9} & -\frac{1}{9} \\ 0 & 1 & 0 & \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array} \right).$$

Finally, we subtract 3 times the second row from the first to eliminate the remaining nonzero off-diagonal entry, thereby completing the Gauss–Jordan procedure:

$$\left( \begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ 0 & 1 & 0 & \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array} \right).$$

The left-hand matrix is the identity, and therefore the final right-hand matrix is our desired inverse:

$$A^{-1} = \left( \begin{array}{ccc} -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array} \right). \quad (1.46)$$

The reader may wish to verify that the final result does satisfy both inverse conditions  $AA^{-1} = I = A^{-1}A$ .

We are now able to complete the proofs of the basic results on inverse matrices. First, we need to determine the elementary matrix corresponding to an elementary row operation of type #3. Again, this is obtained by performing the row operation in question on the identity matrix. Thus, the elementary matrix that multiplies row  $i$  by the nonzero scalar  $c$  is the diagonal matrix having  $c$  in the  $i^{\text{th}}$  diagonal position, and 1's elsewhere along the diagonal. The inverse elementary matrix is the diagonal matrix with  $1/c$  in the  $i^{\text{th}}$  diagonal position and 1's elsewhere on the main diagonal; it corresponds to the inverse operation that divides row  $i$  by  $c$ . For example, the elementary matrix that multiplies the second

row of a 3-rowed matrix by 5 is  $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ ; its inverse is  $E^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

In summary:

**Lemma 1.24.** Every elementary matrix is nonsingular, and its inverse is also an elementary matrix of the same type.

The Gauss–Jordan method tells us how to reduce any nonsingular square matrix  $A$  to the identity matrix by a sequence of elementary row operations. Let  $E_1, E_2, \dots, E_N$  be the corresponding elementary matrices. The elimination procedure that reduces  $A$  to  $I$  amounts to multiplying  $A$  by a succession of elementary matrices:

$$E_N E_{N-1} \cdots E_2 E_1 A = I. \quad (1.47)$$

We claim that the product matrix

$$X = E_N E_{N-1} \cdots E_2 E_1 \quad (1.48)$$

is the inverse of  $A$ . Indeed, formula (1.47) says that  $XA = I$ , and so  $X$  is a left inverse. Furthermore, each elementary matrix has an inverse, and so by (1.41),  $X$  itself is invertible, with

$$X^{-1} = E_1^{-1} E_2^{-1} \cdots E_{N-1}^{-1} E_N^{-1}. \quad (1.49)$$

Therefore, multiplying formula (1.47), namely  $XA = I$ , on the left by  $X^{-1}$  leads to  $A = X^{-1}$ . Lemma 1.20 implies  $X = A^{-1}$ , as claimed, completing the proof of Theorem 1.18. Finally, equating  $A = X^{-1}$  to the product (1.49), and invoking Lemma 1.24, we have established the following result.

**Proposition 1.25.** Every nonsingular matrix can be written as the product of elementary matrices.

**Example 1.26.** The  $2 \times 2$  matrix  $A = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}$  is converted into the identity matrix by first interchanging its rows,  $\begin{pmatrix} 1 & 3 \\ 0 & -1 \end{pmatrix}$ , then scaling the second row by  $-1$ ,  $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$ , and, finally, subtracting 3 times the second row from the first to obtain  $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$ . The corresponding elementary matrices are

$$E_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix}.$$

Therefore, by (1.48),

$$A^{-1} = E_3 E_2 E_1 = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix},$$

while

$$A = E_1^{-1} E_2^{-1} E_3^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}.$$

As an application, let us prove that the inverse of a nonsingular triangular matrix is also triangular. Specifically:

**Proposition 1.27.** If  $L$  is a lower triangular matrix with all nonzero entries on the main diagonal, then  $L$  is nonsingular and its inverse  $L^{-1}$  is also lower triangular. In particular, if  $L$  is lower unitriangular, so is  $L^{-1}$ . A similar result holds for upper triangular matrices.

*Proof:* It suffices to note that if  $L$  has all nonzero diagonal entries, one can reduce  $L$  to the identity by elementary row operations of types #1 and #3, whose associated elementary matrices are all lower triangular. Lemma 1.2 implies that the product (1.48) is then also lower triangular. If  $L$  is unitriangular, then all the pivots are equal to 1. Thus, no elementary row operations of type #3 are required, and so  $L$  can be reduced to the identity matrix by elementary row operations of type #1 alone. Therefore, its inverse is a product of lower unitriangular matrices, and hence is itself lower unitriangular. A similar argument applies in the upper triangular case.  $Q.E.D.$

## Exercises

1.5.24. (a) Write down the elementary matrix that multiplies the third row of a  $4 \times 4$  matrix by 7. (b) Write down its inverse.

1.5.25. Find the inverse of each of the following matrices, if possible, by applying the Gauss–Jordan Method.

$$(a) \begin{pmatrix} 1 & -2 \\ 3 & -3 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} \frac{3}{5} & -\frac{4}{5} \\ \frac{4}{5} & \frac{3}{5} \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 & 0 & -2 \\ 3 & -1 & 0 \\ -2 & 1 & -3 \end{pmatrix},$$

$$(f) \begin{pmatrix} 1 & 2 & 3 \\ 3 & 5 & 5 \\ 2 & 1 & 2 \end{pmatrix}, \quad (g) \begin{pmatrix} 2 & 1 & 2 \\ 4 & 2 & 3 \\ 0 & -1 & 1 \end{pmatrix}, \quad (h) \begin{pmatrix} 2 & 1 & 0 & 1 \\ 0 & 0 & 1 & 3 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & -2 & -5 \end{pmatrix}, \quad (i) \begin{pmatrix} 1 & -2 & 1 & 1 \\ 2 & -3 & 3 & 0 \\ 3 & -7 & 2 & 4 \\ 0 & 2 & 1 & 1 \end{pmatrix}.$$

1.5.26. Write each of the matrices in Exercise 1.5.25 as a product of elementary matrices.

1.5.27. Express  $A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}$  as a product of elementary matrices.

1.5.28. Use the Gauss–Jordan Method to find the inverse of the following complex matrices:

$$(a) \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 1-i \\ 1+i & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 1 & -i \\ i & 0 & -1 \\ -1 & i & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 0 & i \\ i & -1 & 1+i \\ -3i & 1-i & 1+i \end{pmatrix}.$$

1.5.29. Can two nonsingular linear systems have the same solution and yet not be equivalent?

◇ 1.5.30. (a) Suppose  $\tilde{A}$  is obtained from  $A$  by applying an elementary row operation. Let  $C = AB$ , where  $B$  is any matrix of the appropriate size. Explain why  $\tilde{C} = \tilde{A}B$  can be obtained by applying the same elementary row operation to  $C$ . (b) Illustrate by adding

$-2$  times the first row to the third row of  $A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & 2 \\ 0 & 1 & -4 \end{pmatrix}$  and then multiplying the

result on the right by  $B = \begin{pmatrix} 1 & -2 \\ 3 & 0 \\ -1 & 1 \end{pmatrix}$ . Check that the resulting matrix is the same as first

multiplying  $AB$  and then applying the same row operation to the product matrix.

## Solving Linear Systems with the Inverse

The primary motivation for introducing the matrix inverse is that it provides a compact formula for the solution to any linear system with an invertible coefficient matrix.

**Theorem 1.28.** If the matrix  $A$  is nonsingular, then  $\mathbf{x} = A^{-1} \mathbf{b}$  is the unique solution to the linear system  $A \mathbf{x} = \mathbf{b}$ .

*Proof:* We merely multiply the system by  $A^{-1}$ , which yields  $\mathbf{x} = A^{-1}A\mathbf{x} = A^{-1}\mathbf{b}$ . Moreover,  $A\mathbf{x} = AA^{-1}\mathbf{b} = \mathbf{b}$ , proving that  $\mathbf{x} = A^{-1}\mathbf{b}$  is indeed the solution. *Q.E.D.*

For example, let us return to the linear system (1.29). Since we computed the inverse of its coefficient matrix in (1.46), a “direct” way to solve the system is to multiply the right-hand side by the inverse matrix:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{pmatrix} \begin{pmatrix} 2 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{5}{6} \\ \frac{5}{6} \\ \frac{1}{3} \end{pmatrix},$$

reproducing our earlier solution.

However, while aesthetically appealing, the solution method based on the inverse matrix is hopelessly inefficient as compared to direct Gaussian Elimination, and, despite what you may have been told, *should not be used in practical computations*. (A complete justification of this dictum will be provided in Section 1.7.) On the other hand, the inverse does play a useful role in theoretical developments, as well as providing insight into the design of practical algorithms. But the principal message of applied linear algebra is that *LU* decomposition and Gaussian Elimination are fundamental; matrix inverses are to be avoided in all but the most elementary computations.

**Remark.** The reader may have learned a version of the Gauss–Jordan algorithm for solving a single linear system that replaces the Back Substitution step by a complete

reduction of the coefficient matrix to the identity. In other words, to solve  $A\mathbf{x} = \mathbf{b}$ , we start with the augmented matrix  $M = (A \mid \mathbf{b})$  and use all three types of elementary row operations to produce (assuming nonsingularity) the fully reduced form  $(I \mid \mathbf{d})$ , representing the trivially soluble, equivalent system  $\mathbf{x} = \mathbf{d}$ , which is the solution to the original system. However, Back Substitution is more efficient, and it remains the method of choice in practical computations.

## Exercises

1.5.31. Solve the following systems of linear equations by computing the inverses of their coefficient matrices.

$$(a) \begin{array}{l} x + 2y = 1, \\ x - 2y = -2. \end{array} \quad (b) \begin{array}{l} 3u - 2v = 2, \\ u + 5v = 12. \end{array} \quad (c) \begin{array}{l} x - y + 3z = 3, \\ x - 2y + 3z = -2, \\ x - 2y + z = 2. \end{array} \quad (d) \begin{array}{l} y + 5z = 3, \\ x - y + 3z = -1, \\ -2x + 3y = 5. \end{array}$$

$$(e) \begin{array}{l} x + 4y - z = 3, \\ 2x + 7y - 2z = 5, \\ -x - 5y + 2z = -7. \end{array} \quad (f) \begin{array}{l} x + y = 4, \\ 2x + 3y - w = 11, \\ -y - z + w = -7, \\ z - w = 6. \end{array} \quad (g) \begin{array}{l} x - 2y + z + 2u = -2, \\ x - y + z - u = 3, \\ 2x - y + z + u = 3, \\ -x + 3y - 2z - u = 2. \end{array}$$

1.5.32. For each of the nonsingular matrices in Exercise 1.5.25, use your computed inverse to solve the associated linear system  $A\mathbf{x} = \mathbf{b}$ , where  $\mathbf{b}$  is the column vector of the appropriate size that has all 1's as its entries.

## The $LDV$ Factorization

The second phase of the Gauss–Jordan process leads to a slightly more detailed version of the  $LU$  factorization. Let  $D$  denote the diagonal matrix having the same diagonal entries as  $U$ ; in other words,  $D$  contains the pivots on its diagonal and zeros everywhere else. Let  $V$  be the upper unitriangular matrix obtained from  $U$  by dividing each row by its pivot, so that  $V$  has all 1's on the diagonal. We already encountered  $V$  during the course of the Gauss–Jordan procedure. It is easily seen that  $U = DV$ , which implies the following result.

**Theorem 1.29.** A matrix  $A$  is regular if and only if it admits a factorization

$$A = LDV, \tag{1.50}$$

where  $L$  is a lower unitriangular matrix,  $D$  is a diagonal matrix with nonzero diagonal entries, and  $V$  is an upper unitriangular matrix.

For the matrix appearing in Example 1.4, we have  $U = DV$ , where

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad V = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This leads to the factorization

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 5 & 2 \\ 2 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = LDV.$$

**Proposition 1.30.** If  $A = LU$  is regular, then the factors  $L$  and  $U$  are uniquely determined. The same holds for the  $A = LDV$  factorization.

*Proof:* Suppose  $LU = \tilde{L}\tilde{U}$ . Since the diagonal entries of all four matrices are non-zero, Proposition 1.27 implies that they are invertible. Therefore,

$$\tilde{L}^{-1}L = \tilde{L}^{-1}LUU^{-1} = \tilde{L}^{-1}\tilde{L}\tilde{U}U^{-1} = \tilde{U}U^{-1}. \quad (1.51)$$

The left-hand side of the matrix equation (1.51) is the product of two lower unitriangular matrices, and so, by Lemma 1.2, is itself lower unitriangular. The right-hand side is the product of two upper triangular matrices, and hence is upper triangular. But the only way a lower unitriangular matrix can equal an upper triangular matrix is if they both equal the diagonal identity matrix. Therefore,  $\tilde{L}^{-1}L = I = UU^{-1}$ , and so  $\tilde{L} = L$  and  $\tilde{U} = U$ , proving the first result. The  $LDV$  version is an immediate consequence. *Q.E.D.*

As you may have guessed, the more general cases requiring one or more row interchanges lead to a permuted  $LDV$  factorization in the following form.

**Theorem 1.31.** A matrix  $A$  is nonsingular if and only if there is a permutation matrix  $P$  such that

$$PA = LDV, \quad (1.52)$$

where  $L$  is a lower unitriangular matrix,  $D$  is a diagonal matrix with nonzero diagonal entries, and  $V$  is a upper unitriangular matrix.

Uniqueness does not hold for the more general permuted factorizations (1.33), (1.52), since there may be several permutation matrices that place a matrix in regular form; an explicit example can be found in Exercise 1.4.23. Moreover, in contrast to regular Gaussian Elimination, here the pivots, i.e., the diagonal entries of  $U$ , are no longer uniquely defined, but depend on the particular combination of row interchanges employed during the course of the computation.

## Exercises

1.5.33. Produce the  $LDV$  or a permuted  $LDV$  factorization of the following matrices:

$$(a) \begin{pmatrix} 1 & 2 \\ -3 & 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 4 \\ -7 & 2 \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & 1 & 2 \\ 2 & 4 & -1 \\ 0 & -2 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 1 & 5 \\ 1 & 1 & -2 \\ 2 & -1 & 3 \end{pmatrix},$$

$$(e) \begin{pmatrix} 2 & -3 & 2 \\ 1 & -1 & 1 \\ 1 & -1 & 2 \end{pmatrix}, \quad (f) \begin{pmatrix} 1 & -1 & 1 & 2 \\ 1 & -4 & 1 & 5 \\ 1 & 2 & -1 & -1 \\ 3 & 1 & 1 & 6 \end{pmatrix}, \quad (g) \begin{pmatrix} 1 & 0 & 2 & -3 \\ 2 & -2 & 0 & 1 \\ 1 & -2 & -2 & -1 \\ 0 & 1 & 1 & 2 \end{pmatrix}.$$

1.5.34. Using the  $LDV$  factorization for the matrices you found in parts (a–g) of Exercise 1.5.33, solve the corresponding linear systems  $A\mathbf{x} = \mathbf{b}$ , for the indicated vector  $\mathbf{b}$ .

$$(a) \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad (b) \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix}, \quad (d) \begin{pmatrix} -1 \\ 4 \\ -1 \end{pmatrix}, \quad (e) \begin{pmatrix} -1 \\ -2 \\ 5 \end{pmatrix}, \quad (f) \begin{pmatrix} 2 \\ -9 \\ 3 \\ 4 \end{pmatrix}, \quad (g) \begin{pmatrix} 6 \\ -4 \\ 0 \\ -3 \end{pmatrix}.$$

## 1.6 Transposes and Symmetric Matrices

Another basic operation on matrices is to interchange their rows and columns. If  $A$  is an  $m \times n$  matrix, then its *transpose*, denoted by  $A^T$ , is the  $n \times m$  matrix whose  $(i, j)$  entry equals the  $(j, i)$  entry of  $A$ ; thus

$$B = A^T \quad \text{means that} \quad b_{ij} = a_{ji}.$$

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \text{then} \quad A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

Observe that the rows of  $A$  become the columns of  $A^T$  and vice versa. In particular, the transpose of a row vector is a column vector, while the transpose of a column vector is a row vector; if  $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ , then  $\mathbf{v}^T = (1 \ 2 \ 3)$ . The transpose of a scalar, considered as a  $1 \times 1$  matrix, is itself:  $c^T = c$ .

**Remark.** Most vectors appearing in applied mathematics are column vectors. To conserve vertical space in this text, we will often use the transpose notation, e.g.,  $\mathbf{v} = (v_1, v_2, v_3)^T$ , as a compact way of writing the column vector  $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$ .

In the square case, transposition can be viewed as “reflecting” the matrix entries across the main diagonal. For example,

$$\begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 5 \\ -2 & -4 & 8 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & -2 \\ 2 & 0 & -4 \\ -1 & 5 & 8 \end{pmatrix}.$$

In particular, the transpose of a lower triangular matrix is upper triangular and vice-versa.

Transposing twice returns you to where you started:

$$(A^T)^T = A. \quad (1.53)$$

Unlike inversion, transposition *is* compatible with matrix addition and scalar multiplication:

$$(A + B)^T = A^T + B^T, \quad (cA)^T = cA^T. \quad (1.54)$$

Transposition is also compatible with matrix multiplication, but with a twist. Like the inverse, the transpose *reverses* the order of multiplication:

$$(AB)^T = B^T A^T. \quad (1.55)$$

Indeed, if  $A$  has size  $m \times n$  and  $B$  has size  $n \times p$ , so they can be multiplied, then  $A^T$  has size  $n \times m$  and  $B^T$  has size  $p \times n$ , and so, in general, one has no choice but to multiply  $B^T A^T$  in that order. Formula (1.55) is a straightforward consequence of the basic laws of matrix multiplication. More generally,

$$(A_1 A_2 \cdots A_{k-1} A_k)^T = A_k^T A_{k-1}^T \cdots A_2^T A_1^T.$$

An important special case is the product of a row vector  $\mathbf{v}^T$  and a column vector  $\mathbf{w}$  with the same number of entries. In this case,

$$\mathbf{v}^T \mathbf{w} = (\mathbf{v}^T \mathbf{w})^T = \mathbf{w}^T \mathbf{v}, \quad (1.56)$$

because their product is a scalar and so, as noted above, equals its own transpose.

**Lemma 1.32.** If  $A$  is a nonsingular matrix, so is  $A^T$ , and its inverse is denoted by

$$A^{-T} = (A^T)^{-1} = (A^{-1})^T. \quad (1.57)$$

Thus, transposing a matrix and then inverting yields the same result as first inverting and then transposing.

*Proof:* Let  $X = (A^{-1})^T$ . Then, according to (1.55),

$$X A^T = (A^{-1})^T A^T = (A A^{-1})^T = \mathbf{I}^T = \mathbf{I}.$$

The proof that  $A^T X = \mathbf{I}$  is similar, and so we conclude that  $X = (A^T)^{-1}$ . *Q.E.D.*

## Exercises

- 1.6.1. Write down the transpose of the following matrices: (a)  $\begin{pmatrix} 1 \\ 5 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 2 & -1 \\ 2 & 0 & 2 \end{pmatrix}$ , (e)  $\begin{pmatrix} 1 & 2 & -3 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$ , (g)  $\begin{pmatrix} 1 & 2 & -1 \\ 0 & 3 & 2 \\ 1 & 1 & 5 \end{pmatrix}$ .
- 1.6.2. Let  $A = \begin{pmatrix} 3 & -1 & -1 \\ 1 & 2 & 1 \end{pmatrix}$ ,  $B = \begin{pmatrix} -1 & 2 \\ 2 & 0 \\ -3 & 4 \end{pmatrix}$ . Compute  $A^T$  and  $B^T$ . Then compute  $(AB)^T$  and  $(BA)^T$  without first computing  $AB$  or  $BA$ .
- 1.6.3. Show that  $(AB)^T = A^T B^T$  if and only if  $A$  and  $B$  are square commuting matrices.
- ◊ 1.6.4. Prove formula (1.55).
- 1.6.5. Find a formula for the transposed product  $(ABC)^T$  in terms of  $A^T$ ,  $B^T$  and  $C^T$ .
- 1.6.6. *True or false:* Every square matrix  $A$  commutes with its transpose  $A^T$ .
- ◊ 1.6.7. A square matrix is called *normal* if it commutes with its transpose:  $A^T A = A A^T$ . Find all normal  $2 \times 2$  matrices.
- 1.6.8. (a) Prove that the inverse transpose operation (1.57) respects matrix multiplication:  $(AB)^{-T} = A^{-T} B^{-T}$ . (b) Verify this identity for  $A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$ ,  $B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ .
- 1.6.9. Prove that if  $A$  is an invertible matrix, then  $AA^T$  and  $A^T A$  are also invertible.
- 1.6.10. If  $\mathbf{v}$ ,  $\mathbf{w}$  are column vectors with the same number of entries, does  $\mathbf{v} \mathbf{w}^T = \mathbf{w} \mathbf{v}^T$ ?
- 1.6.11. Is there a matrix analogue of formula (1.56), namely  $A^T B = B^T A$ ?
- ◊ 1.6.12. (a) Let  $A$  be an  $m \times n$  matrix. Let  $\mathbf{e}_j$  denote the  $1 \times n$  column vector with a single 1 in the  $j^{\text{th}}$  entry, as in (1.44). Explain why the product  $A \mathbf{e}_j$  equals the  $j^{\text{th}}$  column of  $A$ . (b) Similarly, let  $\hat{\mathbf{e}}_i$  be the  $1 \times m$  column vector with a single 1 in the  $i^{\text{th}}$  entry. Explain why the triple product  $\hat{\mathbf{e}}_i^T A \mathbf{e}_j = a_{ij}$  equals the  $(i, j)$  entry of the matrix  $A$ .

- ◇ 1.6.13. Let  $A$  and  $B$  be  $m \times n$  matrices. (a) Suppose that  $\mathbf{v}^T A \mathbf{w} = \mathbf{v}^T B \mathbf{w}$  for all vectors  $\mathbf{v}, \mathbf{w}$ . Prove that  $A = B$ . (b) Give an example of two matrices such that  $\mathbf{v}^T A \mathbf{v} = \mathbf{v}^T B \mathbf{v}$  for all vectors  $\mathbf{v}$ , but  $A \neq B$ .
- ◇ 1.6.14. (a) Explain why the inverse of a permutation matrix equals its transpose:  $P^{-1} = P^T$ .  
 (b) If  $A^{-1} = A^T$ , is  $A$  necessarily a permutation matrix?
- ◇ 1.6.15. Let  $A$  be a square matrix and  $P$  a permutation matrix of the same size. (a) Explain why the product  $AP^T$  has the effect of applying the permutation defined by  $P$  to the columns of  $A$ . (b) Explain the effect of multiplying  $PAP^T$ . *Hint:* Try this on some  $3 \times 3$  examples first.
- ◇ 1.6.16. Let  $\mathbf{v}, \mathbf{w}$  be  $n \times 1$  column vectors. (a) Prove that in most cases the inverse of the  $n \times n$  matrix  $A = I - \mathbf{v}\mathbf{w}^T$  has the form  $A^{-1} = I - c\mathbf{v}\mathbf{w}^T$  for some scalar  $c$ . Find all  $\mathbf{v}, \mathbf{w}$  for which such a result is valid. (b) Illustrate the method when  $\mathbf{v} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$  and  $\mathbf{w} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ .  
 (c) What happens when the method fails?

## Factorization of Symmetric Matrices

A particularly important class of square matrices consists of those that are unchanged by the transpose operation.

**Definition 1.33.** A matrix is called *symmetric* if it equals its own transpose:  $A = A^T$ .

Thus,  $A$  is symmetric if and only if it is square and its entries satisfy  $a_{ji} = a_{ij}$  for all  $i, j$ . In other words, entries lying in “mirror image” positions relative to the main diagonal must be equal. For example, the most general symmetric  $3 \times 3$  matrix has the form

$$A = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}.$$

Note that all diagonal matrices, including the identity, are symmetric. A lower or upper triangular matrix is symmetric if and only if it is, in fact, a diagonal matrix.

The  $LDV$  factorization of a nonsingular matrix takes a particularly simple form if the matrix also happens to be symmetric. This result will form the foundation of some significant later developments.

**Theorem 1.34.** A symmetric matrix  $A$  is regular if and only if it can be factored as

$$A = LDL^T, \quad (1.58)$$

where  $L$  is a lower unitriangular matrix and  $D$  is a diagonal matrix with nonzero diagonal entries.

*Proof:* We already know, according to Theorem 1.29, that we can factor

$$A = LDV. \quad (1.59)$$

We take the transpose of both sides of this equation:

$$A^T = (LDV)^T = V^T D^T L^T = V^T D L^T, \quad (1.60)$$

since diagonal matrices are automatically symmetric:  $D^T = D$ . Note that  $V^T$  is lower unitriangular, and  $L^T$  is upper unitriangular. Therefore (1.60) is the  $LDV$  factorization of  $A^T$ .

In particular, if  $A$  is symmetric, then

$$LDV = A = A^T = V^T D L^T.$$

Uniqueness of the  $LDV$  factorization implies that

$$L = V^T \quad \text{and} \quad V = L^T$$

(which are two versions of the same equation). Replacing  $V$  by  $L^T$  in (1.59) establishes the factorization (1.58). *Q.E.D.*

**Remark.** If  $A = LDL^T$ , then  $A$  is necessarily symmetric. Indeed,

$$A^T = (LDL^T)^T = (L^T)^T D^T L^T = LDL^T = A.$$

However, not every symmetric matrix has an  $LDL^T$  factorization. A simple example is the irregular but nonsingular  $2 \times 2$  matrix  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ .

**Example 1.35.** The problem is to find the  $LDL^T$  factorization of the particular symmetric matrix  $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$ . This requires performing the usual Gaussian Elimination algorithm. Subtracting twice the first row from the second and also the first row from the third produces the matrix  $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$ . We then add one half of the second row of the latter matrix to its third row, resulting in the upper triangular form

$$U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = DV,$$

which we further factor by dividing each row of  $U$  by its pivot. On the other hand, the lower unitriangular matrix associated with the preceding row operations is  $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix}$ ,

which, as guaranteed by Theorem 1.34, is the transpose of  $V = L^T$ . Therefore, the desired  $A = LU = LDL^T$  factorizations of this particular symmetric matrix are

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = DV.$$

**Example 1.36.** Let us look at a general  $2 \times 2$  symmetric matrix  $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ . Regularity requires that the first pivot be  $a \neq 0$ . A single row operation will place  $A$  in upper triangular form  $U = \begin{pmatrix} a & b \\ 0 & \frac{ac-b^2}{a} \end{pmatrix}$ , and so  $A$  is regular provided  $ac - b^2 \neq 0$ .

also. The associated lower triangular matrix is  $L = \begin{pmatrix} 1 & 0 \\ b & 1 \\ a & \end{pmatrix}$ . Thus,  $A = LU$ , as you can check. Finally,  $D = \begin{pmatrix} a & 0 \\ 0 & \frac{ac-b^2}{a} \end{pmatrix}$  is just the diagonal part of  $U$ , and hence  $U = DL^T$ , so that the  $LDL^T$  factorization is explicitly given by

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ b & 1 \\ \frac{a}{a} & \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & \frac{ac-b^2}{a} \end{pmatrix} \begin{pmatrix} 1 & \frac{b}{a} \\ 0 & 1 \end{pmatrix}. \quad (1.61)$$


---

## Exercises

1.6.17. Find all values of  $a$ ,  $b$ , and  $c$  for which the following matrices are symmetric:

$$(a) \begin{pmatrix} 3 & a \\ 2a-1 & a-2 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & a & 2 \\ -1 & b & c \\ b & 3 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 3 & a+2b-2c & -4 \\ 6 & 7 & b-c \\ -a+b+c & 4 & b+3c \end{pmatrix}.$$

1.6.18. List all symmetric (a)  $3 \times 3$  permutation matrices, (b)  $4 \times 4$  permutation matrices.

1.6.19. *True or false:* If  $A$  is symmetric, then  $A^2$  is symmetric.

◇ 1.6.20. *True or false:* If  $A$  is a nonsingular symmetric matrix, then  $A^{-1}$  is also symmetric.

◇ 1.6.21. *True or false:* If  $A$  and  $B$  are symmetric  $n \times n$  matrices, so is  $AB$ .

1.6.22. (a) Show that every diagonal matrix is symmetric. (b) Show that an upper (lower) triangular matrix is symmetric if and only if it is diagonal.

1.6.23. Let  $A$  be a symmetric matrix. (a) Show that  $A^n$  is symmetric for every nonnegative integer  $n$ . (b) Show that  $2A^2 - 3A + I$  is symmetric. (c) Show that every matrix polynomial  $p(A)$  of  $A$ , cf. Exercise 1.2.35, is a symmetric matrix.

1.6.24. Show that if  $A$  is any matrix, then  $K = A^T A$  and  $L = AA^T$  are both well-defined, symmetric matrices.

1.6.25. Find the  $LDL^T$  factorization of the following symmetric matrices:

$$(a) \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}, \quad (b) \begin{pmatrix} -2 & 3 \\ 3 & -1 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & -1 & -1 \\ -1 & 3 & 2 \\ -1 & 2 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & -1 & 0 & 3 \\ -1 & 2 & 2 & 0 \\ 0 & 2 & -1 & 0 \\ 3 & 0 & 0 & 1 \end{pmatrix}.$$

1.6.26. Find the  $LDL^T$  factorization of the matrices

$$M_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \quad \text{and} \quad M_4 = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}.$$

◇ 1.6.27. Prove that the  $3 \times 3$  matrix  $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & -1 \\ 1 & -1 & 3 \end{pmatrix}$  cannot be factored as  $A = LDL^T$ .

◇ 1.6.28. *Skew-symmetric matrices:* An  $n \times n$  matrix  $J$  is called *skew-symmetric* if  $J^T = -J$ .

- (a) Show that every diagonal entry of a skew-symmetric matrix is zero.
- (b) Write down an example of a nonsingular skew-symmetric matrix.
- (c) Can you find a regular skew-symmetric matrix?
- (d) Show that if  $J$  is a nonsingular skew-symmetric matrix, then  $J^{-1}$  is also skew-symmetric. Verify this fact for the matrix you wrote down in part (b).
- (e) Show that if  $J$  and  $K$  are skew-symmetric, then so are  $J^T$ ,  $J + K$ , and  $J - K$ . What about  $JK$ ?
- (f) Prove that if  $J$  is a skew-symmetric matrix, then  $\mathbf{v}^T J \mathbf{v} = 0$  for all vectors  $\mathbf{v} \in \mathbb{R}^n$ .

1.6.29. (a) Prove that every square matrix can be expressed as the sum,  $A = S + J$ , of a symmetric matrix  $S = S^T$  and a skew-symmetric matrix  $J = -J^T$ .

(b) Write  $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$  and  $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$  as the sum of symmetric and skew-symmetric matrices.

◇ 1.6.30. Suppose  $A = LU$  is a regular matrix. Write down the  $LU$  factorization of  $A^T$ . Prove that  $A^T$  is also regular, and its pivots are the *same* as the pivots of  $A$ .

## 1.7 Practical Linear Algebra

For pedagogical and practical reasons, the examples and exercises we have chosen to illustrate the algorithms are all based on relatively small matrices. When dealing with matrices of moderate size, the differences between the various approaches to solving linear systems (Gauss, Gauss–Jordan, matrix inverse, and so on) are relatively unimportant, particularly if one has a decent computer or even hand calculator to do the tedious parts. However, real-world applied mathematics deals with much larger linear systems, and the design of efficient algorithms is a must. For example, numerical solution schemes for ordinary differential equations will typically lead to matrices with thousands of entries, while numerical schemes for partial differential equations arising in fluid and solid mechanics, weather prediction, image and video processing, quantum mechanics, molecular dynamics, chemical processes, etc., will often require dealing with matrices with more than a million entries. It is not hard for such systems to tax even the most sophisticated supercomputer. Thus, it is essential that we understand the computational details of competing methods in order to compare their efficiency, and thereby gain some experience with the issues underlying the design of high performance numerical algorithms.

The most basic question is this: how many arithmetic operations<sup>†</sup> — in numerical applications these are almost always performed in floating point with various precision levels — are required to complete an algorithm? The number will directly influence the time spent running the algorithm on a computer. We shall keep track of additions and multiplications separately, since the latter typically take longer to process.<sup>‡</sup> But we shall not distinguish between addition and subtraction, nor between multiplication and division, since these typically have the same complexity. We shall also assume that the matrices and vectors we deal with are *generic*, with few, if any, zero entries. Modifications of the basic algorithms for *sparse matrices*, meaning those that have lots of zero entries, are an important topic of research, since these include many of the large matrices that appear in applications to differential equations. We refer the interested reader to more advanced treatments of numerical linear algebra, such as [21, 40, 66, 89], for such developments.

First, when multiplying an  $n \times n$  matrix  $A$  and an  $n \times 1$  column vector  $\mathbf{b}$ , each entry of the product  $A\mathbf{b}$  requires  $n$  multiplications of the form  $a_{ij} b_j$  and  $n - 1$  additions to sum the resulting products. Since there are  $n$  entries, this means a total of  $n^2$  multiplications

<sup>†</sup> For simplicity, we will count only the basic arithmetic operations. But it is worth noting that other issues, such as the number of storage and retrieval operations, may also play a role in estimating the computational complexity of a numerical algorithm.

<sup>‡</sup> At least, in traditional computer architectures. New algorithms and new methods for performing basic arithmetic operations on a computer, particularly in high precision arithmetic, make this discussion trickier. For simplicity, we will stay with the “classical” version here.

and  $n(n - 1) = n^2 - n$  additions. Thus, for a matrix of size  $n = 100$ , one needs about 10,000 distinct multiplications and a similar number of additions. If  $n = 1,000,000 = 10^6$ , then  $n^2 = 10^{12}$ , which is phenomenally large, and the total time required to perform the computation becomes a significant issue<sup>†</sup>.

Let us next look at the (regular) Gaussian Elimination algorithm, referring back to our pseudocode program for the notational details. First, we count how many arithmetic operations are based on the  $j^{\text{th}}$  pivot  $m_{jj}$ . For each of the  $n - j$  rows lying below it, we must perform one division to compute the factor  $l_{ij} = m_{ij}/m_{jj}$  used in the elementary row operation. The entries in the column below the pivot will be set to zero automatically, and so we need only compute the updated entries lying strictly below and to the right of the pivot. There are  $(n - j)^2$  such entries in the coefficient matrix and an additional  $n - j$  entries in the last column of the augmented matrix. Let us concentrate on the former for the moment. For each of these, we replace  $m_{ik}$  by  $m_{ik} - l_{ij}m_{jk}$ , and so must perform one multiplication and one addition. For the  $j^{\text{th}}$  pivot, there is a total of  $(n - j)(n - j + 1)$  multiplications — including the initial  $n - j$  divisions needed to produce the  $l_{ij}$  — and  $(n - j)^2$  additions needed to update the coefficient matrix. Therefore, to reduce a regular  $n \times n$  matrix to upper triangular form requires a total<sup>‡</sup> of

$$\begin{aligned} \sum_{j=1}^n (n - j)(n - j + 1) &= \frac{n^3 - n}{3} && \text{multiplications, and} \\ \sum_{j=1}^n (n - j)^2 &= \frac{2n^3 - 3n^2 + n}{6} && \text{additions.} \end{aligned} \tag{1.62}$$

Thus, when  $n$  is large, both involve approximately  $\frac{1}{3}n^3$  operations.

We should also be keeping track of the number of operations on the right-hand side of the system. No pivots appear there, and so there are

$$\sum_{j=1}^n (n - j) = \frac{n^2 - n}{2} \tag{1.63}$$

multiplications and the same number of additions required to produce the right-hand side in the resulting triangular system  $U\mathbf{x} = \mathbf{c}$ . For large  $n$ , this count is considerably smaller than the coefficient matrix totals (1.62). We note that the Forward Substitution equations (1.26) require precisely the same number of arithmetic operations to solve  $L\mathbf{c} = \mathbf{b}$  for the right-hand side of the upper triangular system. Indeed, the  $j^{\text{th}}$  equation

$$c_j = b_j - \sum_{k=1}^{j-1} l_{jk} c_k$$

requires  $j - 1$  multiplications and the same number of additions, giving a total of

$$\sum_{j=1}^n (j - 1) = \frac{n^2 - n}{2}$$

operations of each type. Therefore, to reduce a linear system to upper triangular form, it makes no difference in computational efficiency whether one works directly with the

<sup>†</sup> See Exercise 1.7.8 for more sophisticated computational algorithms that can be employed to (slightly) speed up multiplication of large matrices.

<sup>‡</sup> In Exercise 1.7.4, the reader is asked to prove these summation formulae by induction.

augmented matrix or employs Forward Substitution after the  $LU$  factorization of the coefficient matrix has been established.

The Back Substitution phase of the algorithm can be similarly analyzed. To find the value of

$$x_j = \frac{1}{u_{jj}} \left( c_j - \sum_{k=j+1}^n u_{jk} x_k \right)$$

once we have computed  $x_{j+1}, \dots, x_n$ , requires  $n - j + 1$  multiplications/divisions and  $n - j$  additions. Therefore, Back Substitution requires

$$\begin{aligned} \sum_{j=1}^n (n - j + 1) &= \frac{n^2 + n}{2} && \text{multiplications, along with} \\ \sum_{j=1}^n (n - j) &= \frac{n^2 - n}{2} && \text{additions.} \end{aligned} \tag{1.64}$$

For  $n$  large, both of these are approximately equal to  $\frac{1}{2}n^2$ . Comparing the counts, we conclude that the bulk of the computational effort goes into the reduction of the coefficient matrix to upper triangular form.

Combining the two counts (1.63–64), we discover that, once we have computed the  $A = LU$  decomposition of the coefficient matrix, the Forward and Back Substitution process requires  $n^2$  multiplications and  $n^2 - n$  additions to solve a linear system  $A\mathbf{x} = \mathbf{b}$ . This is exactly the *same* as the number of multiplications and additions needed to compute the product  $A^{-1}\mathbf{b}$ . Thus, even if we happen to know the inverse of  $A$ , it is still *just as efficient* to use Forward and Back Substitution to compute the solution!

On the other hand, the computation of  $A^{-1}$  is decidedly more inefficient. There are two possible strategies. First, we can solve the  $n$  linear systems (1.45), namely

$$A\mathbf{x} = \mathbf{e}_i, \quad i = 1, \dots, n, \tag{1.65}$$

for the individual columns of  $A^{-1}$ . This requires first computing the  $LU$  decomposition, which uses about  $\frac{1}{3}n^3$  multiplications and a similar number of additions, followed by applying Forward and Back Substitution to each of the systems, using  $n \cdot n^2 = n^3$  multiplications and  $n(n^2 - n) \approx n^3$  additions, for a grand total of about  $\frac{4}{3}n^3$  operations of each type in order to compute  $A^{-1}$ . Gauss–Jordan Elimination fares no better (in fact, slightly worse), also requiring about the same number,  $\frac{4}{3}n^3$ , of each type of arithmetic operation. Both algorithms can be made more efficient by exploiting the fact that there are lots of zeros on the right-hand sides of the systems (1.65). Designing the algorithm to avoid adding or subtracting a preordained 0, or multiplying or dividing by a preordained  $\pm 1$ , reduces the total number of operations required to compute  $A^{-1}$  to exactly  $n^3$  multiplications and  $n(n-1)^2 \approx n^3$  additions. (Details are relegated to the exercises.) And don't forget that we still need to multiply  $A^{-1}\mathbf{b}$  to solve the original system. As a result, solving a linear system with the inverse matrix requires approximately *three* times as many arithmetic operations, and so would take three times as long to complete, as the more elementary Gaussian Elimination and Back Substitution algorithm. This justifies our earlier contention that matrix inversion is inefficient, and, except in very special situations, should never be used for solving linear systems in practice.

## Exercises

1.7.1. Solve the following linear systems by (i) Gaussian Elimination with Back Substitution; (ii) the Gauss–Jordan algorithm to convert the augmented matrix to the fully reduced form  $(I \mid \mathbf{x})$  with solution  $\mathbf{x}$ ; (iii) computing the inverse of the coefficient matrix, and then multiplying it by the right-hand side. Keep track of the number of arithmetic operations you need to perform to complete each computation, and discuss their relative efficiency.

$$(a) \begin{array}{l} x - 2y = 4 \\ 3x + y = -7, \end{array} \quad (b) \begin{array}{l} 2x - 4y + 6z = 6, \\ 3x - 3y + 4z = -1, \\ -4x + 3y - 4z = 5, \end{array} \quad (c) \begin{array}{l} x - 3y = 1, \\ 3x - 7y + 5z = -1, \\ -2x + 6y - 5z = 0. \end{array}$$

1.7.2. (a) Let  $A$  be an  $n \times n$  matrix. Which is faster to compute,  $A^2$  or  $A^{-1}$ ? Justify your answer. (b) What about  $A^3$  versus  $A^{-1}$ ? (c) How many operations are needed to compute  $A^k$ ? Hint: When  $k > 3$ , you can get away with less than  $k - 1$  matrix multiplications!

1.7.3. Which is faster: Back Substitution or multiplying a matrix by a vector? How much faster?

◇ 1.7.4. Use induction to prove the summation formulas (1.62), (1.63) and (1.64).

◇ 1.7.5. Let  $A$  be a general  $n \times n$  matrix. Determine the exact number of arithmetic operations needed to compute  $A^{-1}$  using (a) Gaussian Elimination to factor  $PA = LU$  and then Forward and Back Substitution to solve the  $n$  linear systems (1.65); (b) the Gauss–Jordan method. Make sure your totals do not count adding or subtracting a known 0, or multiplying or dividing by a known  $\pm 1$ .

1.7.6. Count the number of arithmetic operations needed to solve a system the “old-fashioned” way, by using elementary row operations of all three types, in the same order as the Gauss–Jordan scheme, to fully reduce the augmented matrix  $M = (A \mid \mathbf{b})$  to the form  $(I \mid \mathbf{d})$ , with  $\mathbf{x} = \mathbf{d}$  being the solution.

1.7.7. An alternative solution strategy, also called *Gauss–Jordan* in some texts, is, once a pivot is in position, to use elementary row operations of type #1 to eliminate all entries both above and below it, thereby reducing the augmented matrix to diagonal form  $(D \mid \mathbf{c})$  where  $D = \text{diag}(d_1, \dots, d_n)$  is a diagonal matrix containing the pivots. The solutions  $x_i = c_i/d_i$  are then obtained by simple division. Is this strategy more efficient, less efficient, or the same as Gaussian Elimination with Back Substitution? Justify your answer with an exact operations count.

◇ 1.7.8. Here, we describe a remarkable algorithm for matrix multiplication discovered by Strassen, [82]. Let  $A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}$ ,  $B = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}$ , and  $C = \begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix} = AB$  be block matrices of size  $n = 2m$ , where all blocks are of size  $m \times m$ . (a) Let  $D_1 = (A_1 + A_4)(B_1 + B_4)$ ,  $D_2 = (A_1 - A_3)(B_1 + B_2)$ ,  $D_3 = (A_2 - A_4)(B_3 + B_4)$ ,  $D_4 = (A_1 + A_2)B_4$ ,  $D_5 = (A_3 + A_4)B_1$ ,  $D_6 = A_4(B_1 - B_3)$ ,  $D_7 = A_1(B_2 - B_4)$ . Show that  $C_1 = D_1 + D_3 - D_4 - D_6$ ,  $C_2 = D_4 + D_7$ ,  $C_3 = D_5 - D_6$ ,  $C_4 = D_1 - D_2 - D_5 + D_7$ . (b) How many arithmetic operations are required when  $A$  and  $B$  are  $2 \times 2$  matrices? How does this compare with the usual method of multiplying  $2 \times 2$  matrices? (c) In the general case, suppose we use standard matrix multiplication for the matrix products in  $D_1, \dots, D_7$ . Prove that Strassen’s Method is faster than the direct algorithm for computing  $AB$  by a factor of  $\approx \frac{7}{8}$ . (d) When  $A$  and  $B$  have size  $n \times n$  with  $n = 2^r$ , we can recursively apply Strassen’s Method to multiply the  $2^{r-1} \times 2^{r-1}$  blocks  $A_i, B_i$ . Prove that the resulting algorithm requires a total of  $7^r = n^{\log_2 7} = n^{2.80735}$  multiplications

and  $6(7^{r-1} - 4^{r-1}) \leq 7^r = n^{\log_2 7}$  additions/subtractions, versus  $n^3$  multiplications and  $n^3 - n^2 \approx n^3$  additions for the ordinary matrix multiplication algorithm. How much faster is Strassen's Method when  $n = 2^{10}$ ?  $2^{25}$ ?  $2^{100}$ ? (e) How might you proceed if the size of the matrices does not happen to be a power of 2? Further developments of these ideas can be found in [11, 40].

## Tridiagonal Matrices

Of course, in special cases, the actual arithmetic operation count might be considerably reduced, particularly if  $A$  is a sparse matrix with many zero entries. A number of specialized techniques have been designed to handle sparse linear systems. A particularly important class consists of the *tridiagonal matrices*

$$A = \begin{pmatrix} q_1 & r_1 & & & \\ p_1 & q_2 & r_2 & & \\ & p_2 & q_3 & r_3 & \\ & & \ddots & \ddots & \ddots \\ & & & p_{n-2} & q_{n-1} & r_{n-1} \\ & & & & p_{n-1} & q_n \end{pmatrix} \quad (1.66)$$

with all entries zero except for those on the main diagonal, namely  $a_{ii} = q_i$ , the *subdiagonal*, meaning the  $n - 1$  entries  $a_{i+1,i} = p_i$  immediately below the main diagonal, and the *superdiagonal*, meaning the entries  $a_{i,i+1} = r_i$  immediately above the main diagonal. (Blanks are used to indicate 0 entries.) Such matrices arise in the numerical solution of ordinary differential equations and the spline fitting of curves for interpolation and computer graphics. If  $A = LU$  is regular, it turns out that the factors are lower and upper *bidiagonal matrices*, of the form

$$L = \begin{pmatrix} 1 & & & & \\ l_1 & 1 & & & \\ & l_2 & 1 & & \\ & & \ddots & \ddots & \\ & & & l_{n-2} & 1 \\ & & & & l_{n-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} d_1 & u_1 & & & \\ & d_2 & u_2 & & \\ & & d_3 & u_3 & \\ & & & \ddots & \ddots \\ & & & & d_{n-1} & u_{n-1} \\ & & & & & d_n \end{pmatrix}. \quad (1.67)$$

Multiplying out  $LU$  and equating the result to  $A$  leads to the equations

$$\begin{aligned} d_1 &= q_1, & u_1 &= r_1, & l_1 d_1 &= p_1, \\ l_1 u_1 + d_2 &= q_2, & u_2 &= r_2, & l_2 d_2 &= p_2, \\ &\vdots & &\vdots & &\vdots \\ l_{j-1} u_{j-1} + d_j &= q_j, & u_j &= r_j, & l_j d_j &= p_j, \\ &\vdots & &\vdots & &\vdots \\ l_{n-2} u_{n-2} + d_{n-1} &= q_{n-1}, & u_{n-1} &= r_{n-1}, & l_{n-1} d_{n-1} &= p_{n-1}, \\ l_{n-1} u_{n-1} + d_n &= q_n. \end{aligned} \quad (1.68)$$

These elementary algebraic equations can be successively solved for the entries of  $L$  and  $U$  in the following order:  $d_1, u_1, l_1, d_2, u_2, l_2, d_3, u_3, \dots$ . The original matrix  $A$  is regular provided none of the diagonal entries  $d_1, d_2, \dots$  are zero, which allows the recursive procedure to successfully proceed to termination.

Once the  $LU$  factors are in place, we can apply Forward and Back Substitution to solve the tridiagonal linear system  $A\mathbf{x} = \mathbf{b}$ . We first solve the lower triangular system  $L\mathbf{c} = \mathbf{b}$  by Forward Substitution, which leads to the recursive equations

$$c_1 = b_1, \quad c_2 = b_2 - l_1 c_1, \quad \dots \quad c_n = b_n - l_{n-1} c_{n-1}. \quad (1.69)$$

We then solve the upper triangular system  $U\mathbf{x} = \mathbf{c}$  by Back Substitution, again recursively:

$$x_n = \frac{c_n}{d_n}, \quad x_{n-1} = \frac{c_{n-1} - u_{n-1} x_n}{d_{n-1}}, \quad \dots \quad x_1 = \frac{c_1 - u_1 x_2}{d_1}. \quad (1.70)$$

As you can check, there are a total of  $5n - 4$  multiplications/divisions and  $3n - 3$  additions/subtractions required to solve a general tridiagonal system of  $n$  linear equations — a striking improvement over the general case.

**Example 1.37.** Consider the  $n \times n$  tridiagonal matrix

$$A = \begin{pmatrix} 4 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 1 & 4 & 1 \\ & & & & & 1 & 4 \end{pmatrix}$$

in which the diagonal entries are all  $q_i = 4$ , while the entries immediately above and below the main diagonal are all  $p_i = r_i = 1$ . According to (1.68), the tridiagonal factorization (1.67) has  $u_1 = u_2 = \dots = u_{n-1} = 1$ , while

$$d_1 = 4, \quad l_j = 1/d_j, \quad d_{j+1} = 4 - l_j, \quad j = 1, 2, \dots, n-1.$$

The computed values are

$j$	1	2	3	4	5	6	7
$d_j$	4.0	3.75	3.733333	3.732143	3.732057	3.732051	3.732051
$l_j$	.25	.266666	.267857	.267942	.267948	.267949	.267949

These converge rapidly to

$$d_j \rightarrow 2 + \sqrt{3} = 3.732050\dots, \quad l_j \rightarrow 2 - \sqrt{3} = .267949\dots,$$

which makes the factorization for large  $n$  almost trivial. The numbers  $2 \pm \sqrt{3}$  are the roots of the quadratic equation  $x^2 - 4x + 1 = 0$ , and are characterized as the fixed points of the nonlinear iterative system  $d_{j+1} = 4 - 1/d_j$ .

## Exercises

1.7.9. For each of the following tridiagonal systems find the  $LU$  factorization of the coefficient

matrix, and then solve the system. (a)  $\begin{pmatrix} 1 & 2 & 0 \\ -1 & -1 & 1 \\ 0 & -2 & 3 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 4 \\ -1 \\ -6 \end{pmatrix}$ ,

(b)  $\begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & 1 & 0 \\ 0 & -1 & 4 & 1 \\ 0 & 0 & -5 & 6 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 6 \\ 7 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 2 & 0 & 0 \\ -1 & -3 & 0 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & -1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 0 \\ -2 \\ -3 \\ 1 \end{pmatrix}$ .

1.7.10. *True or false:* (a) The product of two tridiagonal matrices is tridiagonal.

(b) The inverse of a tridiagonal matrix is tridiagonal.

1.7.11. (a) Find the  $LU$  factorization of the  $n \times n$  tridiagonal matrix  $A_n$  with all 2's along the diagonal and all -1's along the sub- and super-diagonals for  $n = 3, 4$ , and 5. (b) Use your factorizations to solve the system  $A_n \mathbf{x} = \mathbf{b}$ , where  $\mathbf{b} = (1, 1, 1, \dots, 1)^T$ . (c) Can you write down the  $LU$  factorization of  $A_n$  for general  $n$ ? Do the entries in the factors approach a limit as  $n$  gets larger and larger? (d) Can you find the solution to the system  $A_n \mathbf{x} = \mathbf{b} = (1, 1, 1, \dots, 1)^T$  for general  $n$ ?

♣ 1.7.12. Answer Exercise 1.7.11 if the super-diagonal entries of  $A_n$  are changed to +1.

♣ 1.7.13. Find the  $LU$  factorizations of  $\begin{pmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{pmatrix}$ ,  $\begin{pmatrix} 4 & 1 & 0 & 1 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 1 & 0 & 1 & 4 \end{pmatrix}$ ,  $\begin{pmatrix} 4 & 1 & 0 & 0 & 1 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 1 & 0 & 0 & 1 & 4 \end{pmatrix}$ .

Do you see a pattern? Try the  $6 \times 6$  version. The following exercise should now be clear.

♡ 1.7.14. A *tricirculant matrix*  $C = \begin{pmatrix} q_1 & r_1 & & & & p_1 \\ p_2 & q_2 & r_2 & & & \\ & p_3 & q_3 & r_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & p_{n-1} & q_{n-1} & r_{n-1} \\ & & & p_n & q_n & \end{pmatrix}$  is tridiagonal except

for its  $(1, n)$  and  $(n, 1)$  entries. Tricirculant matrices arise in the numerical solution of periodic boundary value problems and in spline interpolation.

(a) Prove that if  $C = LU$  is regular, its factors have the form

$$\begin{pmatrix} 1 & & & & & \\ l_1 & 1 & & & & \\ l_2 & & 1 & & & \\ & l_3 & & 1 & & \\ & & \ddots & \ddots & & \\ & & & l_{n-2} & 1 & \\ m_1 & m_2 & m_3 & \dots & m_{n-2} & l_{n-1} & 1 \end{pmatrix}, \begin{pmatrix} d_1 & u_1 & & & & v_1 \\ & d_2 & u_2 & & & v_2 \\ & & d_3 & u_3 & & v_3 \\ & & & \ddots & \ddots & \vdots \\ & & & & d_{n-2} & u_{n-2} & v_{n-2} \\ & & & & & d_{n-1} & u_{n-1} & v_{n-1} \\ & & & & & & d_n & \end{pmatrix}.$$

(b) Compute the  $LU$  factorization of the  $n \times n$  tricirculant matrix

$$C_n = \begin{pmatrix} 1 & -1 & & & & -1 \\ -1 & 2 & -1 & & & \\ & -1 & 3 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & n-1 & -1 \\ & & & & -1 & 1 \end{pmatrix} \text{ for } n = 3, 5, \text{ and } 6. \text{ What goes wrong when } n = 4?$$

- ♡ 1.7.15. A matrix  $A$  is said to have *bandwidth*  $k$  if all entries that are more than  $k$  slots away from the main diagonal are zero:  $a_{ij} = 0$  whenever  $|i - j| > k$ . (a) Show that a tridiagonal matrix has band width 1. (b) Write down an example of a  $6 \times 6$  matrix of band width 2 and one of band width 3. (c) Prove that the  $L$  and  $U$  factors of a regular banded matrix have the same band width. (d) Find the  $LU$  factorization of the matrices you wrote down in part (b). (e) Use the factorization to solve the system  $A\mathbf{x} = \mathbf{b}$ , where  $\mathbf{b}$  is the column vector with all entries equal to 1. (f) How many arithmetic operations are needed to solve  $A\mathbf{x} = \mathbf{b}$  if  $A$  is banded? (g) Prove or give a counterexample: the inverse of a banded matrix is banded.

## Pivoting Strategies

Let us now investigate the practical side of pivoting. As we know, in the irregular situations when a zero shows up in a diagonal pivot position, a row interchange is required to proceed with the elimination algorithm. But even when a nonzero pivot element is in place, there may be good numerical reasons for exchanging rows in order to install a more desirable element in the pivot position. Here is a simple example:

$$.01x + 1.6y = 32.1, \quad x + .6y = 22. \quad (1.71)$$

The exact solution to the system is easily found:

$$x = 10, \quad y = 20.$$

Suppose we are working with a very primitive calculator that only retains 3 digits of accuracy. (Of course, this is not a very realistic situation, but the example could be suitably modified to produce similar difficulties no matter how many digits of accuracy our computer is capable of retaining.) The augmented matrix is

$$\left( \begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 1 & .6 & 22 \end{array} \right).$$

Choosing the  $(1, 1)$  entry as our pivot, and subtracting 100 times the first row from the second produces the upper triangular form

$$\left( \begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 0 & -159.4 & -3188 \end{array} \right).$$

Since our calculator has only three-place accuracy, it will round the entries in the second row, producing the augmented coefficient matrix

$$\left( \begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 0 & -159.0 & -3190 \end{array} \right).$$

The solution by Back Substitution gives

$$y = 3190/159 = 20.0628 \dots \simeq 20.1, \quad \text{and then}$$

$$x = 100(32.1 - 1.6y) = 100(32.1 - 32.16) \simeq 100(32.1 - 32.2) = -10.$$

The relatively small error in  $y$  has produced a very large error in  $x$  — not even its sign is correct!

The problem is that the first pivot,  $.01$ , is much smaller than the other element,  $1$ , that appears in the column below it. Interchanging the two rows before performing the row

---



---



---



---



---



---

```

start
  for  $i = 1$  to  $n$ 
    set  $r(i) = i$ 
  next  $i$ 
  for  $j = 1$  to  $n$ 
    if  $m_{r(i),j} = 0$  for all  $i \geq j$ , stop; print “ $A$  is singular”
    choose  $i > j$  such that  $m_{r(i),j}$  is maximal
    interchange  $r(i) \longleftrightarrow r(j)$ 
    for  $i = j + 1$  to  $n$ 
      set  $l_{r(i),j} = m_{r(i),j} / m_{r(j),j}$ 
      for  $k = j + 1$  to  $n + 1$ 
        set  $m_{r(i),k} = m_{r(i),k} - l_{r(i),j} m_{r(j),k}$ 
      next  $k$ 
    next  $i$ 
  next  $j$ 
end

```

---

operation would resolve the difficulty — even with such an inaccurate calculator! After the interchange, we have

$$\left( \begin{array}{cc|c} 1 & .6 & 22 \\ .01 & 1.6 & 32.1 \end{array} \right),$$

which results in the rounded-off upper triangular form

$$\left( \begin{array}{cc|c} 1 & .6 & 22 \\ 0 & 1.594 & 31.88 \end{array} \right) \simeq \left( \begin{array}{cc|c} 1 & .6 & 22 \\ 0 & 1.59 & 31.9 \end{array} \right).$$

The solution by Back Substitution now gives a respectable answer:

$$y = 31.9/1.59 = 20.0628 \dots \simeq 20.1, \quad x = 22 - .6y = 22 - 12.06 \simeq 22 - 12.1 = 9.9.$$

The general strategy, known as *Partial Pivoting*, says that at each stage, we should use the largest (in absolute value) legitimate (i.e., in the pivot column on or below the diagonal) element as the pivot, even if the diagonal element is nonzero. Partial Pivoting can help suppress the undesirable effects of round-off errors during the computation.

In a computer implementation of pivoting, there is no need to waste processor time physically exchanging the row entries in memory. Rather, one introduces a separate array of pointers that serve to indicate which original row is currently in which permuted position. More concretely, one initializes  $n$  row pointers  $r(1) = 1, \dots, r(n) = n$ . Interchanging row  $i$  and row  $j$  of the coefficient or augmented matrix is then accomplished by merely interchanging  $r(i)$  and  $r(j)$ . Thus, to access a matrix element that is currently in row  $i$  of the augmented matrix, one merely retrieves the element that is in row  $r(i)$  in the computer’s memory. An explicit implementation of this strategy is provided in the accompanying pseudocode program.

Partial pivoting will solve most problems, although there can still be difficulties. For instance, it does not accurately solve the system

$$10x + 1600y = 32100, \quad x + .6y = 22,$$

obtained by multiplying the first equation in (1.71) by 1000. The tip-off is that, while the entries in the column containing the pivot are smaller, those in its row are much larger. The solution to this difficulty is *Full Pivoting*, in which one also performs column interchanges — preferably with a column pointer — to move the largest legitimate element into the pivot position. In practice, a column interchange amounts to reordering the variables in the system, which, as long as one keeps proper track of the order, also doesn't change the solutions. Thus, switching the order of  $x, y$  leads to the augmented matrix

$$\left( \begin{array}{cc|c} 1600 & 10 & 32100 \\ .6 & 1 & 22 \end{array} \right),$$

in which the first column now refers to  $y$  and the second to  $x$ . Now Gaussian Elimination will produce a reasonably accurate solution to the system.

Finally, there are some matrices that are hard to handle even with sophisticated pivoting strategies. Such *ill-conditioned* matrices are typically characterized by being “almost” singular. A famous example of an ill-conditioned matrix is the  $n \times n$  *Hilbert matrix*

$$H_n = \left( \begin{array}{cccccc} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots & \frac{1}{n+2} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \cdots & \frac{1}{n+3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \frac{1}{n+3} & \cdots & \frac{1}{2n-1} \end{array} \right). \quad (1.72)$$

Later, in Proposition 3.40, we will prove that  $H_n$  is nonsingular for all  $n$ . However, the solution of a linear system whose coefficient matrix is a Hilbert matrix  $H_n$ , even for moderately large  $n$ , is a very challenging problem, even using high precision computer arithmetic<sup>†</sup>. This is because the larger  $n$  is, the closer  $H_n$  is, in a sense, to being singular. A full discussion of the so-called condition number of a matrix can be found in Section 8.7.

The reader is urged to try the following computer experiment. Fix a moderately large value of  $n$ , say 20. Choose a column vector  $\mathbf{x}$  with  $n$  entries chosen at random. Compute  $\mathbf{b} = H_n \mathbf{x}$  directly. Then try to solve the system  $H_n \mathbf{x} = \mathbf{b}$  by Gaussian Elimination, and compare the result with the original vector  $\mathbf{x}$ . If you obtain an accurate solution with  $n = 20$ , try  $n = 50$  or 100. This will give you a good indicator of the degree of arithmetic precision used by your computer hardware, and the accuracy of the numerical solution algorithm(s) in your software.

---

<sup>†</sup> In computer algebra systems such as MAPLE and MATHEMATICA, one can use exact rational arithmetic to perform the computations. Then the important issues are time and computational efficiency. Incidentally, there is an explicit formula for the inverse of a Hilbert matrix, which appears in Exercise 1.7.23.

## Exercises

1.7.16. (a) Find the exact solution to the linear system  $\begin{pmatrix} .1 & 2.7 \\ 1.0 & .5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 10. \\ -6.0 \end{pmatrix}$ . (b) Solve

the system using Gaussian Elimination with 2-digit rounding. (c) Solve the system using Partial Pivoting and 2-digit rounding. (d) Compare your answers and discuss.

1.7.17. (a) Find the exact solution to the linear system  $x - 5y - z = 1$ ,  $\frac{1}{6}x - \frac{5}{6}y + z = 0$ ,  $2x - y = 3$ . (b) Solve the system using Gaussian Elimination with 4-digit rounding.

(c) Solve the system using Partial Pivoting and 4-digit rounding. Compare your answers.

1.7.18. Answer Exercise 1.7.17 for the system

$$x + 4y - 3z = -3, \quad 25x + 97y - 35z = 39, \quad 35x - 22y + 33z = -15.$$

1.7.19. Employ 2 digit arithmetic with rounding to compute an approximate solution of the linear system  $0.2x + 2y - 3z = 6$ ,  $5x + 43y + 27z = 58$ ,  $3x + 23y - 42z = -87$ , using the following methods: (a) Regular Gaussian Elimination with Back Substitution; (b) Gaussian Elimination with Partial Pivoting; (c) Gaussian Elimination with Full Pivoting. (d) Compare your answers and discuss their accuracy.

1.7.20. Solve the following systems by hand, using pointers instead of physically interchanging

the rows: (a)  $\begin{pmatrix} 0 & 1 & -2 \\ 1 & -1 & 1 \\ 3 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 0 & -1 & 0 & -1 \\ 0 & 0 & -2 & 1 \\ 1 & 0 & 2 & 0 \\ -1 & 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ ,

(c)  $\begin{pmatrix} 3 & -1 & 2 & -1 \\ 6 & -2 & 4 & 3 \\ 3 & 1 & 0 & -2 \\ -1 & 3 & -2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 0 & -1 & 5 & -1 \\ 1 & -2 & 0 & 1 \\ 2 & -3 & -3 & -1 \\ 2 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 3 \\ 0 \end{pmatrix}$ .

1.7.21. Solve the following systems using Partial Pivoting and pointers:

(a)  $\begin{pmatrix} 1 & 5 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 2 & -1 \\ 4 & -2 & 1 \\ 3 & 5 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$ ,

(c)  $\begin{pmatrix} 1 & -3 & 6 & -1 \\ 2 & -5 & 0 & 1 \\ -1 & -6 & 4 & -2 \\ 3 & 0 & 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 0 \\ 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} .01 & 4 & 2 \\ 2 & -802 & 3 \\ 7 & .03 & 250 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 122 \end{pmatrix}$ .

1.7.22. Use Full Pivoting with pointers to solve the systems in Exercise 1.7.21.

♣ 1.7.23. Let  $H_n$  be the  $n \times n$  Hilbert matrix (1.72), and  $K_n = H_n^{-1}$  its inverse. It can be proved, [40; p. 513], that the  $(i, j)$  entry of  $K_n$  is

$$(-1)^{i+j}(i+j-1) \binom{n+i-1}{n-j} \binom{n+j-1}{n-i} \binom{i+j-2}{i-1}^2,$$

where  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  is the standard binomial coefficient. (Warning. Proving this formula is a nontrivial combinatorial challenge.) (a) Write down the inverse of the Hilbert matrices  $H_3, H_4, H_5$  using the formula or the Gauss–Jordan Method with exact rational arithmetic. Check your results by multiplying the matrix by its inverse.

(b) Recompute the inverses on your computer using floating point arithmetic and compare with the exact answers. (c) Try using floating point arithmetic to find  $K_{10}$  and  $K_{20}$ . Test the answer by multiplying the Hilbert matrix by its computed inverse.

♣ 1.7.24. (a) Write out a pseudo-code algorithm, using both row and column pointers, for Gaussian Elimination with Full Pivoting. (b) Implement your code on a computer, and try it on the systems in Exercise 1.7.21.

## 1.8 General Linear Systems

So far, we have treated only linear systems involving the same number of equations as unknowns, and then only those with nonsingular coefficient matrices. These are precisely the systems that always have a unique solution. We now turn to the problem of solving a general linear system of  $m$  equations in  $n$  unknowns. The cases not treated as yet are non-square systems, with  $m \neq n$ , as well as square systems with singular coefficient matrices. The basic idea underlying the Gaussian Elimination algorithm for nonsingular systems can be straightforwardly adapted to these cases, too. One systematically applies the same two types of elementary row operation to reduce the coefficient matrix to a simplified form that generalizes the upper triangular form we aimed for in the nonsingular situation.

**Definition 1.38.** An  $m \times n$  matrix  $U$  is said to be in *row echelon form* if it has the following “staircase” structure:

$$U = \left( \begin{array}{cccc|cccccccc|cccccccc} \textcircled{*} & * & \dots & * & * & \dots & * & * & \dots & \dots & * & * & * & \dots & * \\ 0 & 0 & \dots & 0 & \textcircled{*} & \dots & * & * & \dots & \dots & * & * & * & \dots & * \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \textcircled{*} & \dots & \dots & * & * & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & \textcircled{*} & * & \dots & * \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \end{array} \right)$$

The entries indicated by  $\textcircled{*}$  are the *pivots*, and must be nonzero. The first  $r$  rows of  $U$  each contain exactly one pivot, but not all columns are required to include a pivot entry. The entries below the “staircase”, indicated by the solid line, are all zero, while the non-pivot entries above the staircase, indicated by stars, can be anything. The last  $m - r$  rows are identically zero, and do not contain any pivots. There may, in exceptional situations, be one or more all zero initial columns. Here is an explicit example of a matrix in row echelon form:

$$\begin{pmatrix} 3 & 1 & 0 & 4 & 5 & -7 \\ 0 & -1 & -2 & 1 & 8 & 0 \\ 0 & 0 & 0 & 0 & 2 & -4 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The three pivots are the first nonzero entries in the three nonzero rows, namely, 3,  $-1$ , 2.

Slightly more generally,  $U$  may have several initial columns consisting of all zeros. An example is the row echelon matrix

$$\begin{pmatrix} 0 & 0 & 3 & 5 & -2 & 0 \\ 0 & 0 & 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 0 & 0 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

which also has three pivots. The latter matrix corresponds to a linear system in which the first two variables do not appear in any of the equations. Thus, such row echelon forms almost never appear in applications.

**Proposition 1.39.** Every matrix can be reduced to row echelon form by a sequence of elementary row operations of types #1 and #2.

In matrix language, Proposition 1.39 implies that if  $A$  is any  $m \times n$  matrix, then there exists an  $m \times m$  permutation matrix  $P$  and an  $m \times m$  lower unitriangular matrix  $L$  such that

$$PA = LU, \quad (1.73)$$

where  $U$  is an  $m \times n$  row echelon matrix. The factorization (1.73) is not unique. Observe that  $P$  and  $L$  are square matrices of the same size, while  $A$  and  $U$  are rectangular, also of the same size. As with a square matrix, the entries of  $L$  below the diagonal correspond to the row operations of type #1, while  $P$  keeps track of row interchanges. As before, one can keep track of row interchanges with a row pointer.

A constructive proof of this result is based on the general Gaussian Elimination algorithm, which proceeds as follows. Starting on the left of the matrix, one searches for the first column that is not identically zero. Any of the nonzero entries in that column may serve as the pivot. Partial pivoting indicates that it is probably best to choose the largest one, although this is not essential for the algorithm to proceed. One places the chosen pivot in the first row of the matrix via a row interchange, if necessary. The entries below the pivot are made equal to zero by the appropriate elementary row operations of type #1. One then proceeds iteratively, performing the same reduction algorithm on the submatrix consisting of all entries strictly to the right and below the pivot. The algorithm terminates when either there is a nonzero pivot in the last row, or all of the rows lying below the last pivot are identically zero, and so no more pivots can be found.

**Example 1.40.** The easiest way to learn the general Gaussian Elimination algorithm is to follow through an illustrative example. Consider the linear system

$$\begin{aligned} x + 3y + 2z - u &= a, \\ 2x + 6y + z + 4u + 3v &= b, \\ -x - 3y - 3z + 3u + v &= c, \\ 3x + 9y + 8z - 7u + 2v &= d, \end{aligned} \quad (1.74)$$

of 4 equations in 5 unknowns, where  $a, b, c, d$  are given numbers<sup>†</sup>. The coefficient matrix is

$$A = \begin{pmatrix} 1 & 3 & 2 & -1 & 0 \\ 2 & 6 & 1 & 4 & 3 \\ -1 & -3 & -3 & 3 & 1 \\ 3 & 9 & 8 & -7 & 2 \end{pmatrix}. \quad (1.75)$$

To solve the system, we introduce the augmented matrix

$$\left( \begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 2 & 6 & 1 & 4 & 3 & b \\ -1 & -3 & -3 & 3 & 1 & c \\ 3 & 9 & 8 & -7 & 2 & d \end{array} \right),$$

obtained by appending the right-hand side of the system. The upper left entry is nonzero, and so can serve as the first pivot. We eliminate the entries below it by elementary row

---

<sup>†</sup> It will be convenient to work with the right-hand side in general form, although the reader may prefer, at least initially, to assign numerical values to  $a, b, c, d$ .

operations, resulting in

$$\left( \begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b - 2a \\ 0 & 0 & -1 & 2 & 1 & c + a \\ 0 & 0 & 2 & -4 & 2 & d - 3a \end{array} \right).$$

Now, the second column contains no suitable nonzero entry to serve as the second pivot. (The top entry already lies in a row containing a pivot, and so cannot be used.) Therefore, we move on to the third column, choosing the  $(2, 3)$  entry,  $-3$ , as our second pivot. Again, we eliminate the entries below it, leading to

$$\left( \begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b - 2a \\ 0 & 0 & 0 & 0 & 0 & c - \frac{1}{3}b + \frac{5}{3}a \\ 0 & 0 & 0 & 0 & 4 & d + \frac{2}{3}b - \frac{13}{3}a \end{array} \right).$$

The fourth column has no pivot candidates, and so the final pivot is the  $4$  in the fifth column. We interchange the last two rows in order to place the coefficient matrix in row echelon form:

$$\left( \begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b - 2a \\ 0 & 0 & 0 & 0 & 4 & d + \frac{2}{3}b - \frac{13}{3}a \\ 0 & 0 & 0 & 0 & 0 & c - \frac{1}{3}b + \frac{5}{3}a \end{array} \right). \quad (1.76)$$

There are three pivots,  $1, -3$ , and  $4$ , sitting in positions  $(1, 1)$ ,  $(2, 3)$ , and  $(3, 5)$ . Note the staircase form, with the pivots on the steps and everything below the staircase being zero. Recalling the row operations used to construct the solution (and keeping in mind that the row interchange that appears at the end also affects the entries of  $L$ ), we find the factorization (1.73) takes the explicit form

$$\left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array} \right) \left( \begin{array}{ccccc} 1 & 3 & 2 & -1 & 0 \\ 2 & 6 & 1 & 4 & 3 \\ -1 & -3 & -3 & 3 & 1 \\ 3 & 9 & 8 & -7 & 2 \end{array} \right) = \left( \begin{array}{cccc} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -\frac{2}{3} & 1 & 0 \\ -1 & \frac{1}{3} & 0 & 1 \end{array} \right) \left( \begin{array}{ccccc} 1 & 3 & 2 & -1 & 0 \\ 0 & 0 & -3 & 6 & 3 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

We shall return to find the solution to our linear system after a brief theoretical interlude.

**Warning.** In the augmented matrix, pivots can *never* appear in the last column, representing the right-hand side of the system. Thus, even if  $c - \frac{1}{3}b + \frac{5}{3}a \neq 0$ , that entry does not qualify as a pivot.

We now introduce the most important numerical quantity associated with a matrix.

**Definition 1.41.** The *rank* of a matrix is the number of pivots.

For instance, the rank of the matrix (1.75) equals  $3$ , since its reduced row echelon form, i.e., the first five columns of (1.76), has three pivots. Since there is at most one pivot per row and one pivot per column, the rank of an  $m \times n$  matrix is bounded by both  $m$  and  $n$ , and so

$$0 \leq r = \text{rank } A \leq \min\{m, n\}. \quad (1.77)$$

The only  $m \times n$  matrix of rank  $0$  is the zero matrix  $\mathbf{0}$  — which is the only matrix without any pivots.

**Proposition 1.42.** A square matrix of size  $n \times n$  is nonsingular if and only if its rank is equal to  $n$ .

Indeed, the only way an  $n \times n$  matrix can end up having  $n$  pivots is if its reduced row echelon form is upper triangular with nonzero diagonal entries. But a matrix that reduces to such triangular form is, by definition, nonsingular.

Interestingly, the rank of a matrix *does not depend* on which elementary row operations are performed along the way to row echelon form. Indeed, performing a different sequence of row operations — say using Partial Pivoting versus no pivoting — can produce a completely different reduced form. The remarkable result is that all such row echelon forms end up having exactly the same number of pivots, and this number is the rank of the matrix. A formal proof of this fact will appear in Chapter 2; see Theorem 2.49.

Once the coefficient matrix has been reduced to row echelon form ( $U | \mathbf{c}$ ), the solution to the equivalent linear system  $U \mathbf{x} = \mathbf{c}$  proceeds as follows. The first step is to see whether there are any equations that do not have a solution. Suppose one of the rows in the echelon form  $U$  is identically zero, but the corresponding entry in the last column  $\mathbf{c}$  of the augmented matrix is nonzero. What linear equation would this represent? Well, the coefficients of all the variables are zero, and so the equation is of the form

$$0 = c_i, \quad (1.78)$$

where  $i$  is the row's index. If  $c_i \neq 0$ , then the equation cannot be satisfied — it is *inconsistent*. The reduced system does not have a solution. Since the reduced system was obtained by elementary row operations, the original linear system is *incompatible*, meaning it also has no solutions. *Note:* It takes only one inconsistency to render the entire system incompatible. On the other hand, if  $c_i = 0$ , so the entire row in the augmented matrix is zero, then (1.78) is merely  $0 = 0$ , and is trivially satisfied. Such all-zero rows do not affect the solvability of the system.

In our example, the last row in the echelon form (1.76) is all zero, and hence the last entry in the final column must also vanish in order that the system be compatible. Therefore, the linear system (1.74) will have a solution if and only if the right-hand sides  $a, b, c, d$  satisfy the linear constraint

$$\frac{5}{3}a - \frac{1}{3}b + c = 0. \quad (1.79)$$

In general, if the system is incompatible, there is nothing else to do. Otherwise, every all zero row in the row echelon form of the coefficient matrix also has a zero entry in the last column of the augmented matrix; the system is *compatible* and admits one or more solutions. (If there are no all-zero rows in the coefficient matrix, meaning that every row contains a pivot, then the system is automatically compatible.) To find the solution(s), we split the variables in the system into two classes.

**Definition 1.43.** In a linear system  $U \mathbf{x} = \mathbf{c}$  in row echelon form, the variables corresponding to columns containing a pivot are called *basic variables*, while the variables corresponding to the columns without a pivot are called *free variables*.

The solution to the system then proceeds by an adaptation of the Back Substitution procedure. Working in reverse order, each nonzero equation is solved for the basic variable associated with its pivot. The result is substituted into the preceding equations before they in turn are solved. The solution then specifies all the basic variables as certain combinations of the remaining free variables. As their name indicates, the free variables, if

any, are allowed to take on any values whatsoever, and so serve to parameterize the general solution to the system.

**Example 1.44.** Let us illustrate the solution procedure with our particular system (1.74). The values  $a = 0, b = 3, c = 1, d = 1$ , satisfy the consistency constraint (1.79), and the corresponding reduced augmented matrix (1.76) is

$$\left( \begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & 0 \\ 0 & 0 & -3 & 6 & 3 & 3 \\ 0 & 0 & 0 & 0 & 4 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right).$$

The pivots are found in columns 1, 3, 5, and so the corresponding variables,  $x, z, v$ , are basic; the other variables,  $y, u$ , corresponding to the non-pivot columns 2, 4, are free. Our task is to solve the reduced system

$$\begin{aligned} x + 3y + 2z - u &= 0, \\ -3z + 6u + 3v &= 3, \\ 4v &= 3, \\ 0 &= 0, \end{aligned}$$

for the basic variables  $x, z, v$  in terms of the free variables  $y, u$ . As before, this is done in the reverse order, by successively substituting the resulting values in the preceding equation. The result is the general solution

$$v = \frac{3}{4}, \quad z = -1 + 2u + v = -\frac{1}{4} + 2u, \quad x = -3y - 2z + u = \frac{1}{2} - 3y - 3u.$$

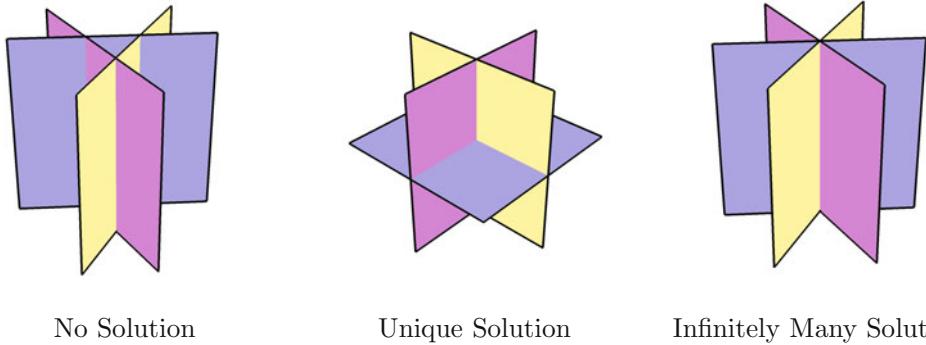
The free variables  $y, u$  remain completely arbitrary; any assigned values will produce a solution to the original system. For instance, if  $y = -1, u = \pi$ , then  $x = \frac{7}{2} - 3\pi$ ,  $z = -\frac{1}{4} + 2\pi$ ,  $v = \frac{3}{4}$ . But keep in mind that this is merely one of an infinite number of valid solutions.

In general, if the  $m \times n$  coefficient matrix of a system of  $m$  linear equations in  $n$  unknowns has rank  $r$ , there are  $m - r$  all-zero rows in the row echelon form, and these  $m - r$  equations must have zero right-hand side in order that the system be compatible and have a solution. Moreover, there is a total of  $r$  basic variables and  $n - r$  free variables, and so the general solution depends upon  $n - r$  parameters.

Summarizing the preceding discussion, we have learned that there are only three possible outcomes for the solution to a system of linear equations.

**Theorem 1.45.** A system  $A\mathbf{x} = \mathbf{b}$  of  $m$  linear equations in  $n$  unknowns has either (i) exactly one solution, (ii) infinitely many solutions, or (iii) no solution.

Case (iii) occurs if the system is incompatible, producing a zero row in the echelon form that has a nonzero right-hand side. Case (ii) occurs if the system is compatible and there are one or more free variables, and so the rank of the coefficient matrix is strictly less than the number of columns:  $r < n$ . Case (i) occurs for nonsingular square coefficient matrices, and, more generally, for compatible systems for which  $r = n$ , implying there are no free variables. Since  $r \leq m$ , this case can arise only if the coefficient matrix has at least as many rows as columns, i.e., the linear system has at least as many equations as unknowns. A linear system can *never* have a finite number — other than 0 or 1 — of solutions. As a consequence, any linear system that admits two or more solutions automatically has infinitely many!



**Figure 1.1.** Intersecting Planes.

**Warning.** This property requires linearity, and is *not* valid for nonlinear systems. For instance, the real quadratic equation  $x^2 + x - 2 = 0$  has exactly two real solutions:  $x = 1$  and  $x = -2$ .

**Example 1.46.** Consider the linear system

$$y + 4z = a, \quad 3x - y + 2z = b, \quad x + y + 6z = c,$$

consisting of three equations in three unknowns. The augmented coefficient matrix is

$$\left( \begin{array}{ccc|c} 0 & 1 & 4 & a \\ 3 & -1 & 2 & b \\ 1 & 1 & 6 & c \end{array} \right).$$

Interchanging the first two rows, and then eliminating the elements below the first pivot leads to

$$\left( \begin{array}{ccc|c} 3 & -1 & 2 & b \\ 0 & 1 & 4 & a \\ 0 & \frac{4}{3} & \frac{16}{3} & c - \frac{1}{3}b \end{array} \right).$$

The second pivot is in the  $(2, 2)$  position, but after eliminating the entry below it, we find the row echelon form to be

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

Since there is a row of all zeros, the original coefficient matrix is singular, and its rank is only 2.

The consistency condition follows from this last row in the reduced echelon form, which requires

$$\frac{4}{3}a + \frac{1}{3}b - c = 0.$$

If this is not satisfied, the system has no solutions; otherwise, it has infinitely many. The free variable is  $z$ , since there is no pivot in the third column. The general solution is

$$y = a - 4z, \quad x = \frac{1}{3}b + \frac{1}{3}y - \frac{2}{3}z = \frac{1}{3}a + \frac{1}{3}b - 2z,$$

where  $z$  is arbitrary.

Geometrically, Theorem 1.45 is telling us about the possible configurations of linear subsets (lines, planes, etc.) of an  $n$ -dimensional space. For example, a single linear equation  $ax + by + cz = d$ , with  $(a, b, c) \neq \mathbf{0}$ , defines a plane  $P$  in three-dimensional space. The solutions to a system of three linear equations in three unknowns belong to all three planes; that is, they lie in their *intersection*  $P_1 \cap P_2 \cap P_3$ . Generically, three planes intersect in a single common point; this is case (i) of the theorem, which occurs if and only if the coefficient matrix is nonsingular. The case of infinitely many solutions occurs when the three planes intersect in a common line, or, even more degenerately, when they all coincide. On the other hand, parallel planes, or planes intersecting in parallel lines, have no common point of intersection, and this occurs when the system is incompatible and has no solutions. There are no other possibilities: the total number of points in the intersection is either 0, 1, or  $\infty$ . Some sample geometric configurations appear in Figure 1.1.

## Exercises

1.8.1. Which of the following systems has (i) a unique solution? (ii) infinitely many

solutions? (iii) no solution? In each case, find all solutions: (a)  $\begin{array}{l} x - 2y = 1, \\ 3x + 2y = -3. \end{array}$

$$(b) \begin{array}{l} 2x + y + 3z = 1, \\ x + 4y - 2z = -3. \end{array} \quad (c) \begin{array}{l} x + y - 2z = -3, \\ 2x - y + 3z = 7, \\ x - 2y + 5z = 1. \end{array} \quad (d) \begin{array}{l} x - 2y + z = 6, \\ 2x + y - 3z = -3, \\ x - 3y + 3z = 10. \end{array}$$

$$(e) \begin{array}{l} x - 2y + 2z - w = 3, \\ 3x + y + 6z + 11w = 16, \\ 2x - y + 4z + w = 9. \end{array} \quad (f) \begin{array}{l} 3x - 2y + z = 4, \\ x + 3y - 4z = -3, \\ 2x - 3y + 5z = 7, \\ x - 8y + 9z = 10. \end{array} \quad (g) \begin{array}{l} x + 2y + 17z - 5w = 50, \\ 9x - 16y + 10z - 8w = 24, \\ 2x - 5y - 4z = -13, \\ 6x - 12y + z - 4w = -1. \end{array}$$

1.8.2. Determine if the following systems are compatible and, if so, find the general solution:

$$(a) \begin{array}{l} 6x_1 + 3x_2 = 12, \\ 4x_1 + 2x_2 = 9. \end{array} \quad (b) \begin{array}{l} 8x_1 + 12x_2 = 16, \\ 6x_1 + 9x_2 = 13. \end{array} \quad (c) \begin{array}{l} x_1 + 2x_2 = 1, \\ 2x_1 + 5x_2 = 2, \\ 3x_1 + 6x_2 = 3. \end{array} \quad (d) \begin{array}{l} 2x_1 - 6x_2 + 4x_3 = 2, \\ -x_1 + 3x_2 - 2x_3 = -1. \end{array}$$

$$(e) \begin{array}{l} 2x_1 + 2x_2 + 3x_3 = 1, \\ x_2 + 2x_3 = 3, \\ 4x_1 + 5x_2 + 7x_3 = 15. \end{array} \quad (f) \begin{array}{l} x_1 + x_2 + x_3 + 9x_4 = 8, \\ x_2 + 2x_3 + 8x_4 = 7, \\ -3x_1 + x_3 - 7x_4 = 9. \end{array} \quad (g) \begin{array}{l} x_1 + 2x_2 + 3x_3 + 4x_4 = 1, \\ 2x_1 + 4x_2 + 6x_3 + 5x_4 = 0, \\ 3x_1 + 4x_2 + x_3 + x_4 = 0, \\ 4x_1 + 6x_2 + 4x_3 - x_4 = 0. \end{array}$$

1.8.3. Graph the following planes and determine whether they have a common intersection:

$$x + y + z = 1, \quad x + y = 1, \quad x + z = 1.$$

1.8.4. Let  $A = \left( \begin{array}{ccc|c} a & 0 & b & 2 \\ a & 2 & a & b \\ b & 2 & a & a \end{array} \right)$  be the augmented matrix for a linear system. For which

values of  $a$  and  $b$  does the system have (i) a unique solution? (ii) infinitely many solutions? (iii) no solution?

1.8.5. Determine the general (complex) solution to the following systems:

$$(a) \begin{array}{l} 2x + (1 + i)y - 2iz = 2i, \\ (1 - i)x + y - 2iz = 0. \end{array} \quad (b) \begin{array}{l} x + 2iy + (2 - 4i)z = 5 + 5i, \\ (-1 + i)x + 2y + (4 + 2i)z = 0, \\ (1 - i)x + (1 + 4i)y - 5iz = 10 + 5i. \end{array}$$

$$\begin{array}{ll}
 x_1 + ix_2 + x_3 = 1 + 4i, & (2+i)x + iy + (2+2i)z + (1+12i)w = 0, \\
 (c) -x_1 + x_2 - ix_3 = -1, & (d) (1-i)x + y + (2-i)z + (8+2i)w = 0, \\
 ix_1 - x_2 - x_3 = -1 - 2i. & (3+2i)x + iy + (3+3i)z + 19iw = 0.
 \end{array}$$

1.8.6. For which values of  $b$  and  $c$  does the system  $x_1 + x_2 + bx_3 = 1$ ,  $bx_1 + 3x_2 - x_3 = -2$ ,  $3x_1 + 4x_2 + x_3 = c$ , have (a) no solution? (b) exactly one solution? (c) infinitely many solutions?

1.8.7. Determine the rank of the following matrices: (a)  $\begin{pmatrix} 1 & -1 \\ 1 & -2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 2 & 1 & 3 \\ -2 & -1 & -3 \end{pmatrix}$ ,  
 (c)  $\begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 2 \\ -1 & 1 & 0 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & -1 & 0 \\ 2 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$ , (e)  $\begin{pmatrix} 3 \\ 0 \\ -2 \end{pmatrix}$ , (f)  $\begin{pmatrix} 0 & -1 & 2 & 5 \end{pmatrix}$ ,  
 (g)  $\begin{pmatrix} 0 & -3 \\ 4 & -1 \\ 1 & 2 \\ -1 & -5 \end{pmatrix}$ , (h)  $\begin{pmatrix} 1 & -1 & 2 & 1 \\ 2 & 1 & -1 & 0 \\ 1 & 2 & -3 & -1 \\ 4 & -1 & 3 & 2 \\ 0 & 3 & -5 & -2 \end{pmatrix}$ , (i)  $\begin{pmatrix} 0 & 0 & 0 & 3 & 1 \\ 1 & 2 & -3 & 1 & -2 \\ 2 & 4 & -2 & 1 & -2 \end{pmatrix}$ .

1.8.8. Write out a  $PA = LU$  factorization for each of the matrices in Exercise 1.8.7.

1.8.9. Construct a system of three linear equations in three unknowns that has  
 (a) one and only one solution; (b) more than one solution; (c) no solution.

1.8.10. Find a coefficient matrix  $A$  such that the associated linear system  $A\mathbf{x} = \mathbf{b}$  has  
 (a) infinitely many solutions for every  $\mathbf{b}$ ; (b) 0 or  $\infty$  solutions, depending on  $\mathbf{b}$ ;  
 (c) 0 or 1 solution depending on  $\mathbf{b}$ ; (d) exactly 1 solution for all  $\mathbf{b}$ .

1.8.11. Give an example of a *nonlinear* system of two equations in two unknowns that has  
 (a) no solution; (b) exactly two solutions; (c) exactly three solutions; (d) infinitely many solutions.

1.8.12. What does it mean if a linear system has a coefficient matrix with a column of all 0's?

1.8.13. *True or false:* One can find an  $m \times n$  matrix of rank  $r$  for every  $0 \leq r \leq \min\{m, n\}$ .

1.8.14. *True or false:* Every  $m \times n$  matrix has (a) exactly  $m$  pivots; (b) at least one pivot.

◇ 1.8.15. (a) Prove that the product  $A = \mathbf{v}\mathbf{w}^T$  of a nonzero  $m \times 1$  column vector  $\mathbf{v}$  and a nonzero  $1 \times n$  row vector  $\mathbf{w}^T$  is an  $m \times n$  matrix of rank  $r = 1$ . (b) Compute the following rank one products: (i)  $\begin{pmatrix} 1 \\ 3 \end{pmatrix} \begin{pmatrix} -1 & 2 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 4 \\ 0 \\ -2 \end{pmatrix} \begin{pmatrix} -2 & 1 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 2 \\ -3 \end{pmatrix} \begin{pmatrix} 1 & 3 & -1 \end{pmatrix}$ .

(c) Prove that every rank one matrix can be written in the form  $A = \mathbf{v}\mathbf{w}^T$ .

◇ 1.8.16. (a) Let  $A$  be an  $m \times n$  matrix and let  $M = (A \mid \mathbf{b})$  be the augmented matrix for the linear system  $A\mathbf{x} = \mathbf{b}$ . Show that either (i)  $\text{rank } A = \text{rank } M$ , or (ii)  $\text{rank } A = \text{rank } M - 1$ .  
 (b) Prove that the system is compatible if and only if case (i) holds.

1.8.17. Find the rank of the matrix  $\begin{pmatrix} a & ar & \dots & ar^{n-1} \\ ar^n & ar^{n+1} & \dots & ar^{2n-1} \\ \vdots & \vdots & \ddots & \vdots \\ ar^{(n-1)n} & ar^{(n-1)n+1} & \dots & ar^{n^2-1} \end{pmatrix}$  when  $a, r \neq 0$ .

1.8.18. Find the rank of the  $n \times n$  matrix  $\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ n+1 & n+2 & n+3 & \dots & 2n \\ 2n+1 & 2n+2 & 2n+3 & \dots & 3n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n^2 - n + 1 & n^2 - n + 2 & \dots & \dots & n^2 \end{pmatrix}$ .

1.8.19. Find two matrices  $A, B$  such that  $\text{rank } AB \neq \text{rank } BA$ .

- ◇ 1.8.20. Let  $A$  be an  $m \times n$  matrix of rank  $r$ . (a) Suppose  $C = (A \ B)$  is an  $m \times k$  matrix,  $k > n$ , whose first  $n$  columns are the same as the columns of  $A$ . Prove that  $\text{rank } C \geq \text{rank } A$ . Give an example with  $\text{rank } C = \text{rank } A$ ; with  $\text{rank } C > \text{rank } A$ . (b) Let  $E = \begin{pmatrix} A \\ D \end{pmatrix}$  be a  $j \times n$  matrix,  $j > m$ , whose first  $m$  rows are the same as those of  $A$ . Prove that  $\text{rank } E \geq \text{rank } A$ . Give an example with  $\text{rank } E = \text{rank } A$ ; with  $\text{rank } E > \text{rank } A$ .
- ◇ 1.8.21. Let  $A$  be a singular square matrix. Prove that there exist elementary matrices  $E_1, \dots, E_N$  such that  $A = E_1 E_2 \cdots E_N Z$ , where  $Z$  is a matrix with at least one all-zero row.

## Homogeneous Systems

A linear system with all 0's on the right-hand side is called *homogeneous*. Conversely, if at least one of the right-hand sides is nonzero, the system is called *inhomogeneous*.

In matrix notation, a homogeneous system takes the form

$$A \mathbf{x} = \mathbf{0}, \quad (1.80)$$

where the zero vector  $\mathbf{0}$  indicates that every entry on the right-hand side is zero. Homogeneous systems are always compatible, since  $\mathbf{x} = \mathbf{0}$  is a solution, known as the *trivial solution*. If a homogeneous system has a nontrivial solution  $\mathbf{x} \neq \mathbf{0}$ , then Theorem 1.45 assures us that it must have infinitely many solutions. This will occur if and only if the reduced system has one or more free variables.

**Theorem 1.47.** A homogeneous linear system  $A \mathbf{x} = \mathbf{0}$  of  $m$  equations in  $n$  unknowns has a nontrivial solution  $\mathbf{x} \neq \mathbf{0}$  if and only if the rank of  $A$  is  $r < n$ . If  $m < n$ , the system *always* has a nontrivial solution. If  $m = n$ , the system has a nontrivial solution if and only if  $A$  is singular.

Thus, homogeneous systems with fewer equations than unknowns always have infinitely many solutions. Indeed, the coefficient matrix of such a system has more columns than rows, and so at least one column cannot contain a pivot, meaning that there is at least one free variable in the general solution formula.

**Example 1.48.** Consider the homogeneous linear system

$$\begin{aligned} 2x_1 + x_2 + 5x_4 &= 0, \\ 4x_1 + 2x_2 - x_3 + 8x_4 &= 0, \\ -2x_1 - x_2 + 3x_3 - 4x_4 &= 0, \end{aligned}$$

with coefficient matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 5 \\ 4 & 2 & -1 & 8 \\ -2 & -1 & 3 & -4 \end{pmatrix}.$$

Since there are only three equations in four unknowns, we already know that the system has infinitely many solutions, including the trivial solution  $x_1 = x_2 = x_3 = x_4 = 0$ .

When solving a homogeneous system, the final column of the augmented matrix consists of all zeros. As such, it will never be altered by row operations, and so it is a waste of effort to carry it along during the process. We therefore perform the Gaussian Elimination

algorithm directly on the coefficient matrix  $A$ . Working with the  $(1, 1)$  entry as the first pivot, we first obtain

$$\begin{pmatrix} 2 & 1 & 0 & 5 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 3 & 1 \end{pmatrix}.$$

The  $(2, 3)$  entry is the second pivot, and we apply one final row operation to place the matrix in row echelon form

$$\begin{pmatrix} 2 & 1 & 0 & 5 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & -5 \end{pmatrix}.$$

This corresponds to the reduced homogeneous system

$$2x_1 + x_2 + 5x_4 = 0, \quad -x_3 - 2x_4 = 0, \quad -5x_4 = 0.$$

Since there are three pivots in the final row echelon form, the rank of the coefficient matrix  $A$  is 3. There is one free variable, namely  $x_2$ . Using Back Substitution, we easily obtain the general solution

$$x_1 = -\frac{1}{2}t, \quad x_2 = t, \quad x_3 = x_4 = 0,$$

which depends upon a single free parameter  $t = x_2$ .

**Example 1.49.** Consider the homogeneous linear system

$$\begin{aligned} 2x - y + 3z &= 0, \\ -4x + 2y - 6z &= 0, \\ 2x - y + z &= 0, \\ 6x - 3y + 3z &= 0, \end{aligned} \quad \text{with coefficient matrix} \quad A = \begin{pmatrix} 2 & -1 & 3 \\ -4 & 2 & -6 \\ 2 & -1 & 1 \\ 6 & -3 & 3 \end{pmatrix}.$$

The system admits the trivial solution  $x = y = z = 0$ , but in this case we need to complete the elimination algorithm before we know for sure whether there are other solutions. After

the first stage in the reduction process, the coefficient matrix becomes

$$\begin{pmatrix} 2 & -1 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & -6 \end{pmatrix}.$$

To continue, we need to interchange the second and third rows to place a nonzero entry in

the final pivot position; after that, the reduction to the row echelon form

$$\begin{pmatrix} 2 & -1 & 3 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is immediate. Thus, the system reduces to the equations

$$2x - y + 3z = 0, \quad -2z = 0, \quad 0 = 0, \quad 0 = 0.$$

The third and fourth equations are trivially compatible, as they must be in the homogeneous case. The rank of the coefficient matrix is equal to two, which is less than the number of columns, and so, even though the system has more equations than unknowns, it has infinitely many solutions. These can be written in terms of the free variable  $y$ ; the general solution is  $x = \frac{1}{2}y$ ,  $z = 0$ , where  $y$  is arbitrary.

---

## Exercises

1.8.22. Solve the following homogeneous linear systems.

$$(a) \begin{array}{l} x + y - 2z = 0, \\ -x + 4y - 3z = 0. \end{array} \quad (b) \begin{array}{l} 2x + 3y - z = 0, \\ -4x + 3y - 5z = 0, \\ x - 3y + 3z = 0. \end{array} \quad (c) \begin{array}{l} -x + y - 4z = 0, \\ -2x + 2y - 6z = 0, \\ x + 3y + 3z = 0. \end{array}$$

$$(d) \begin{array}{l} x + 2y - 2z + w = 0, \\ -3x + z - 2w = 0. \end{array} \quad (e) \begin{array}{l} -x + 3y - 2z + w = 0, \\ -2x + 5y + z - 2w = 0, \\ 3x - 8y + z - 4w = 0. \end{array} \quad (f) \begin{array}{l} -y + z = 0, \\ 2x - 3w = 0, \\ x + y - 2w = 0, \\ y - 3z + w = 0. \end{array}$$

1.8.23. Find all solutions to the homogeneous system  $A\mathbf{x} = \mathbf{0}$  for the coefficient matrix

$$(a) \begin{pmatrix} 3 & -1 \\ -9 & 3 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & -1 & 4 \\ 3 & 1 & 2 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & -2 & 3 & -3 \\ 2 & 1 & 4 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix},$$

$$(e) \begin{pmatrix} 0 & 2 & -1 \\ -2 & 0 & 3 \\ 1 & 3 & 0 \end{pmatrix}, \quad (f) \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 2 & -1 \\ 1 & 0 \end{pmatrix}, \quad (g) \begin{pmatrix} 1 & 2 & 0 \\ -1 & -3 & 2 \\ 4 & 7 & 2 \\ -1 & 1 & 6 \end{pmatrix}, \quad (h) \begin{pmatrix} 0 & 0 & 3 & -3 \\ 1 & -1 & 0 & 3 \\ 2 & -2 & 1 & 5 \\ -1 & 1 & 1 & -4 \end{pmatrix}.$$

1.8.24. Let  $U$  be an upper triangular matrix. Show that the homogeneous system  $U\mathbf{x} = \mathbf{0}$  admits a nontrivial solution if and only if  $U$  has at least one 0 on its diagonal.

1.8.25. Find the solution to the homogeneous system  $2x_1 + x_2 - 2x_3 = 0$ ,  $2x_1 - x_2 - 2x_3 = 0$ . Then solve the inhomogeneous version where the right-hand sides are changed to  $a, b$ , respectively. What do you observe?

1.8.26. Answer Exercise 1.8.25 for the system  $2x_1 + x_2 + x_3 - x_4 = 0$ ,  $2x_1 - 2x_2 - x_3 + 3x_4 = 0$ .

1.8.27. Find all values of  $k$  for which the following homogeneous systems of linear equations have a non-trivial solution:

$$(a) \begin{array}{l} x + ky = 0, \\ kx + 4y = 0, \end{array} \quad (b) \begin{array}{l} x_1 + kx_2 + 4x_3 = 0, \\ kx_1 + x_2 + 2x_3 = 0, \\ 2x_1 + kx_2 + 8x_3 = 0. \end{array} \quad (c) \begin{array}{l} x + ky + 2z = 0, \\ 3x - ky - 2z = 0, \\ (k+1)x - 2y - 4z = 0, \\ kx + 3y + 6z = 0. \end{array}$$

---

## 1.9 Determinants

You may be surprised that, so far, we have not mentioned determinants — a topic that typically assumes a central role in many treatments of basic linear algebra. Determinants can be useful in low-dimensional and highly structured problems, and have many fascinating properties. They also prominently feature in theoretical developments of the subject. But, like matrix inverses, they are almost completely irrelevant when it comes to large scale applications and practical computations. Indeed, for most matrices, the best way to compute a determinant is (surprise) Gaussian Elimination! Consequently, from a computational standpoint, the determinant adds no new information concerning the linear system and its solutions. However, for completeness and in preparation for certain later developments (particularly computing eigenvalues of small matrices), you should be familiar with the basic facts and properties of determinants, as summarized in this final section.

The determinant of a square matrix<sup>†</sup>  $A$  is a scalar, written  $\det A$ , that will distinguish between singular and nonsingular matrices. We already encountered in (1.38) the determinant of a  $2 \times 2$  matrix<sup>‡</sup>:  $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$ . The key fact is that the determinant is nonzero if and only if the matrix has an inverse, or, equivalently, is nonsingular. Our goal is to find an analogous quantity for general square matrices.

There are many different ways to define determinants. The difficulty is that the actual formula is very unwieldy — see (1.87) below — and not well motivated. We prefer an axiomatic approach that explains how our three elementary row operations affect the determinant.

**Theorem 1.50.** Associated with every square matrix, there exists a uniquely defined scalar quantity, known as its *determinant*, that obeys the following axioms:

- (i) Adding a multiple of one row to another does not change the determinant.
- (ii) Interchanging two rows changes the sign of the determinant.
- (iii) Multiplying a row by any scalar (including zero) multiplies the determinant by the same scalar.
- (iv) The determinant of an upper triangular matrix  $U$  is equal to the product of its diagonal entries:  $\det U = u_{11}u_{22} \cdots u_{nn}$ .

In particular, axiom (iv) implies that the determinant of the identity matrix is

$$\det I = 1. \quad (1.81)$$

Checking that all four of these axioms hold in the  $2 \times 2$  case is an elementary exercise.

The proof of Theorem 1.50 is based on the following results. Suppose, in particular, we multiply a row of the matrix  $A$  by the zero scalar. The resulting matrix has a row of all zeros, and, by axiom (iii), has zero determinant. Since any matrix with a zero row can be obtained in this fashion, we conclude:

**Lemma 1.51.** Any matrix with one or more all-zero rows has zero determinant.

Using these properties, one is able to compute the determinant of any square matrix by Gaussian Elimination, which is, in fact, the fastest and most practical computational method in all but the simplest situations.

**Theorem 1.52.** If  $A = LU$  is a regular matrix, then

$$\det A = \det U = u_{11}u_{22} \cdots u_{nn} \quad (1.82)$$

equals the product of the pivots. More generally, if  $A$  is nonsingular, and requires  $k$  row interchanges to arrive at its permuted factorization  $PA = LU$ , then

$$\det A = \det P \det U = (-1)^k u_{11}u_{22} \cdots u_{nn}. \quad (1.83)$$

Finally,  $A$  is singular if and only if

$$\det A = 0. \quad (1.84)$$

---

<sup>†</sup> Non-square matrices do not have determinants.

<sup>‡</sup> Some authors use vertical lines to indicate the determinant:  $\left| \begin{matrix} a & b \\ c & d \end{matrix} \right| = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ .

*Proof:* In the regular case, we need only elementary row operations of type #1 to reduce  $A$  to upper triangular form  $U$ , and axiom (i) says these do not change the determinant. Therefore,  $\det A = \det U$ , the formula for the latter being given by axiom (iv). The nonsingular case follows in a similar fashion. By axiom (ii), each row interchange changes the sign of the determinant, and so  $\det A$  equals  $\det U$  if there has been an even number of interchanges, but equals  $-\det U$  if there has been an odd number. For the same reason, the determinant of the permutation matrix  $P$  equals  $+1$  if there has been an even number of row interchanges, and  $-1$  for an odd number. Finally, if  $A$  is singular, then we can reduce it to a matrix with at least one row of zeros by elementary row operations of types #1 and #2. Lemma 1.51 implies that the resulting matrix has zero determinant, and so  $\det A = 0$ , also.  $Q.E.D.$

**Remark.** If we then apply Gauss–Jordan elimination to reduce the upper triangular matrix  $U$  to the identity matrix  $I$ , and use axiom (ii) when each row is divided by its pivot, we find that axiom (iv) follows from the simpler formula (1.81), which could thus replace it in Theorem 1.50.

**Example 1.53.** Let us compute the determinant of the  $4 \times 4$  matrix

$$A = \begin{pmatrix} 1 & 0 & -1 & 2 \\ 2 & 1 & -3 & 4 \\ 0 & 2 & -2 & 3 \\ 1 & 1 & -4 & -2 \end{pmatrix}.$$

We perform our usual Gaussian Elimination algorithm, successively leading to the matrices

$$A \mapsto \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 2 & -2 & 3 \\ 0 & 1 & -3 & -4 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & -2 & -4 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & 0 & 3 \end{pmatrix},$$

where we used a single row interchange to obtain the final upper triangular form. Owing to the row interchange, the determinant of the original matrix is  $-1$  times the product of the pivots:

$$\det A = -1 \cdot (1 \cdot 1 \cdot (-2) \cdot 3) = 6.$$

In particular, this tells us that  $A$  is nonsingular. But, of course, this was already evident, since we successfully reduced the matrix to upper triangular form with 4 nonzero pivots.

There is a variety of other approaches to evaluating determinants. However, except for very small ( $2 \times 2$  or  $3 \times 3$ ) matrices or other special situations, the most efficient algorithm for computing the determinant of a matrix is to apply Gaussian Elimination, with pivoting if necessary, and then invoke the relevant formula from Theorem 1.52. In particular, the determinantal criterion (1.84) for singular matrices, while of theoretical interest, is unnecessary in practice, since we will have already detected whether the matrix is singular during the course of the elimination procedure by observing that it has fewer than the full number of pivots.

Let us finish by stating a few of the basic properties of determinants. Proofs are outlined in the exercises.

**Proposition 1.54.** The determinant of the product of two square matrices of the same size is the product of their determinants:

$$\det(AB) = \det A \det B. \quad (1.85)$$

Therefore, even though matrix multiplication is not commutative, and so  $AB \neq BA$  in general, both matrix products have the same determinant:

$$\det(AB) = \det A \det B = \det B \det A = \det(BA),$$

because ordinary (scalar) multiplication *is* commutative. In particular, setting  $B = A^{-1}$  and using axiom (iv), we find that the determinant of the inverse matrix is the reciprocal of the matrix's determinant.

**Proposition 1.55.** If  $A$  is a nonsingular matrix, then

$$\det A^{-1} = \frac{1}{\det A}. \quad (1.86)$$

Finally, for later reference, we end with the general formula for the determinant of an  $n \times n$  matrix  $A$  with entries  $a_{ij}$ :

$$\det A = \sum_{\pi} (\text{sign } \pi) a_{\pi(1),1} a_{\pi(2),2} \cdots a_{\pi(n),n}. \quad (1.87)$$

The sum is over all possible permutations  $\pi$  of the rows of  $A$ . The *sign* of the permutation, written  $\text{sign } \pi$ , equals the determinant of the corresponding permutation matrix  $P$ , so  $\text{sign } \pi = \det P = +1$  if the permutation is composed of an even number of row interchanges and  $-1$  if composed of an odd number. For example, the six terms in the well-known formula

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}a_{22}a_{33} + a_{31}a_{12}a_{23} + a_{21}a_{32}a_{13} - a_{11}a_{32}a_{23} - a_{21}a_{12}a_{33} - a_{31}a_{22}a_{13} \quad (1.88)$$

for a  $3 \times 3$  determinant correspond to the six possible permutations (1.31) of a 3-rowed matrix. A proof that the formula (1.87) satisfies the defining properties of the determinant listed in Theorem 1.50 is tedious, but not hard. The reader might wish to try out the  $3 \times 3$  case to be convinced that it works.

The explicit formula (1.87) proves that the determinant function is well-defined, and formally completes the proof of Theorem 1.50. One consequence of this formula is that the determinant is unaffected by the transpose operation.

**Proposition 1.56.** Transposing a matrix does not change its determinant:

$$\det A^T = \det A. \quad (1.89)$$

**Remark.** Proposition 1.56 has the interesting consequence that one can equally well use “elementary column operations” to compute determinants. We will not develop this approach in any detail here, since it does not help us to solve linear equations.

However, the explicit determinant formula (1.87) is not used in practice. Since there are  $n!$  different permutations of the  $n$  rows, the determinantal sum (1.87) contains  $n!$  distinct terms, which, as soon as  $n$  is of moderate size, renders it completely useless for practical computations. For instance, the determinant of a  $10 \times 10$  matrix contains  $10! = 3,628,800$

terms, while a  $100 \times 100$  determinant would require summing  $9.3326 \times 10^{157}$  terms, each of which is a product of 100 matrix entries! The most efficient way to compute determinants is still our mainstay — Gaussian Elimination, coupled with the fact that the determinant is  $\pm$  the product of the pivots! On this note, we conclude our brief introduction.

## Exercises

1.9.1. Use Gaussian Elimination to find the determinant of the following matrices:

$$(a) \begin{pmatrix} 2 & -1 \\ -4 & 3 \end{pmatrix}, (b) \begin{pmatrix} 0 & 1 & -2 \\ -1 & 0 & 3 \\ 2 & -3 & 0 \end{pmatrix}, (c) \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 10 \end{pmatrix}, (d) \begin{pmatrix} 0 & 1 & -1 \\ -2 & 1 & 3 \\ 2 & 7 & -8 \end{pmatrix},$$

$$(e) \begin{pmatrix} 5 & -1 & 0 & 2 \\ 0 & 3 & -1 & 5 \\ 0 & 0 & -4 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}, (f) \begin{pmatrix} 1 & -2 & 1 & 4 \\ 2 & -4 & 0 & 0 \\ 3 & -4 & 2 & 5 \\ 0 & 2 & -4 & -9 \end{pmatrix}, (g) \begin{pmatrix} 1 & -2 & 1 & 4 & -5 \\ 1 & 1 & -2 & 3 & -3 \\ 2 & -1 & -1 & 2 & 2 \\ 5 & -1 & 0 & 5 & 5 \\ 2 & 2 & 0 & 4 & -1 \end{pmatrix}.$$

1.9.2. Verify the determinant product formula (1.85) when

$$A = \begin{pmatrix} 1 & -1 & 3 \\ 2 & -1 & 1 \\ 4 & -2 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & -1 \\ 1 & -3 & -2 \\ 2 & 0 & 1 \end{pmatrix}.$$

1.9.3. (a) Give an example of a non-diagonal  $2 \times 2$  matrix for which  $A^2 = I$ . (b) In general, if  $A^2 = I$ , show that  $\det A = \pm 1$ . (c) If  $A^2 = A$ , what can you say about  $\det A$ ?

1.9.4. *True or false:* If true, explain why. If false, give an explicit counterexample.

- (a) If  $\det A \neq 0$  then  $A^{-1}$  exists. (b)  $\det(2A) = 2 \det A$ . (c)  $\det(A + B) = \det A + \det B$ .
- (d)  $\det A^{-T} = \frac{1}{\det A}$ . (e)  $\det(AB^{-1}) = \frac{\det A}{\det B}$ . (f)  $\det[(A + B)(A - B)] = \det(A^2 - B^2)$ .
- (g) If  $A$  is an  $n \times n$  matrix with  $\det A = 0$ , then  $\text{rank } A < n$ .
- (h) If  $\det A = 1$  and  $AB = O$ , then  $B = O$ .

1.9.5. Prove that the similar matrices  $B = S^{-1}AS$  have the same determinant:  $\det A = \det B$ .

1.9.6. Prove that if  $A$  is a  $n \times n$  matrix and  $c$  is a scalar, then  $\det(cA) = c^n \det A$ .

1.9.7. Prove that the determinant of a lower triangular matrix is the product of its diagonal entries.

1.9.8. (a) Show that if  $A$  has size  $n \times n$ , then  $\det(-A) = (-1)^n \det A$ . (b) Prove that, for  $n$  odd, any  $n \times n$  skew-symmetric matrix  $A = -A^T$  is singular. (c) Find a nonsingular skew-symmetric matrix.

◇ 1.9.9. Prove directly that the  $2 \times 2$  determinant formula (1.38) satisfies the four determinant axioms listed in Theorem 1.50.

◇ 1.9.10. In this exercise, we prove the determinantal product formula (1.85). (a) Prove that if  $E$  is any elementary matrix (of the appropriate size), then  $\det(EB) = \det E \det B$ . (b) Use induction to prove that if  $A = E_1 E_2 \cdots E_N$  is a product of elementary matrices, then  $\det(AB) = \det A \det B$ . Explain why this proves the product formula whenever  $A$  is a nonsingular matrix. (c) Prove that if  $Z$  is a matrix with a zero row, then  $ZB$  also has a zero row, and so  $\det(ZB) = 0 = \det Z \det B$ . (d) Use Exercise 1.8.21 to complete the proof of the product formula.

1.9.11. Prove (1.86).

◇ 1.9.12. Prove (1.89). *Hint:* Use Exercise 1.6.30 in the regular case. Then extend to the nonsingular case. Finally, explain why the result also holds for singular matrices.

1.9.13. Write out the formula for a  $4 \times 4$  determinant. It should contain  $24 = 4!$  terms.

◇ 1.9.14. Show that (1.87) satisfies all four determinant axioms, and hence is the correct formula for a determinant.

◇ 1.9.15. Prove that axiom (iv) in Theorem 1.50 can be proved as a consequence of the first three axioms and the property  $\det I = 1$ .

◇ 1.9.16. Prove that one cannot produce an elementary row operation of type #2 by a combination of elementary row operations of type #1.

◇ 1.9.17. Show that (a) if  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is regular, then its pivots are  $a$  and  $\frac{\det A}{a}$ ;

(b) if  $A = \begin{pmatrix} a & b & e \\ c & d & f \\ g & h & j \end{pmatrix}$  is regular, then its pivots are  $a, \frac{ad-bc}{a}$ , and  $\frac{\det A}{ad-bc}$ .

(c) Can you generalize this observation to regular  $n \times n$  matrices?

◇ 1.9.18. In this exercise, we justify the use of “elementary column operations” to compute determinants. Prove that (a) adding a scalar multiple of one column to another does not change the determinant; (b) multiplying a column by a scalar multiplies the determinant by the same scalar; (c) interchanging two columns changes the sign of the determinant.

(d) Explain how to use elementary column operations to reduce a matrix to lower triangular form and thereby compute its determinant.

◇ 1.9.19. Find the determinant of the Vandermonde matrices listed in Exercise 1.3.24. Can you guess the general  $n \times n$  formula?

◇ 1.9.20. *Cramer’s Rule.* (a) Show that the nonsingular system  $ax + by = p, cx + dy = q$  has the solution given by the determinantal ratios

$$x = \frac{1}{\Delta} \det \begin{pmatrix} p & b \\ q & d \end{pmatrix}, \quad y = \frac{1}{\Delta} \det \begin{pmatrix} a & p \\ c & q \end{pmatrix}, \quad \text{where } \Delta = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix}. \quad (1.90)$$

(b) Use Cramer’s Rule (1.90) to solve the systems (i)  $\begin{array}{l} x + 3y = 13, \\ 4x + 2y = 0, \end{array}$  (ii)  $\begin{array}{l} x - 2y = 4, \\ 3x + 6y = -2. \end{array}$

$$ax + by + cz = p,$$

(c) Prove that the solution to  $\begin{array}{l} dx + ey + fz = q, \\ gx + hy + jz = r, \end{array}$  with  $\Delta = \det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & j \end{pmatrix} \neq 0$  is

$$x = \frac{1}{\Delta} \det \begin{pmatrix} p & b & c \\ q & e & f \\ r & h & j \end{pmatrix}, \quad y = \frac{1}{\Delta} \det \begin{pmatrix} a & p & c \\ d & q & f \\ g & r & j \end{pmatrix}, \quad z = \frac{1}{\Delta} \det \begin{pmatrix} a & b & p \\ d & e & q \\ g & h & r \end{pmatrix}. \quad (1.91)$$

$$x + 4y = 3, \quad 3x + 2y - z = 1,$$

(d) Use Cramer’s Rule (1.91) to solve (i)  $4x + 2y + z = 2, \quad (ii) \begin{array}{l} x - 3y + 2z = 2, \\ -x + y - z = 0, \end{array}$  (iii)  $2x - y + z = 3.$

(e) Can you see the pattern that will generalize to  $n$  equations in  $n$  unknowns?

**Remark.** Although elegant, Cramer’s rule is not a very practical solution method.

◇ 1.9.21. (a) Show that if  $D = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$  is a block diagonal matrix, where  $A$  and  $B$  are square matrices, then  $\det D = \det A \det B$ . (b) Prove that the same holds for a block upper

triangular matrix  $\det \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} = \det A \det B$ . (c) Use this method to compute the determinant of the following matrices:

$$(i) \begin{pmatrix} 3 & 2 & -2 \\ 0 & 4 & -5 \\ 0 & 3 & 7 \end{pmatrix}, (ii) \begin{pmatrix} 1 & 2 & -2 & 5 \\ -3 & 1 & 0 & -5 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 2 & -2 \end{pmatrix}, (iii) \begin{pmatrix} 1 & 2 & 0 & 4 \\ -3 & 1 & 4 & -1 \\ 0 & 3 & 1 & 8 \\ 0 & 0 & 0 & -3 \end{pmatrix}, (iv) \begin{pmatrix} 5 & -1 & 0 & 0 \\ 2 & 5 & 0 & 0 \\ 2 & 4 & 4 & -2 \\ 3 & -2 & 9 & -5 \end{pmatrix}.$$



# Chapter 2

## Vector Spaces and Bases

Vector spaces and their ancillary structures provide the common language of linear algebra, and, as such, are an essential prerequisite for understanding contemporary applied (and pure) mathematics. The key concepts of vector space, subspace, linear independence, span, and basis will appear, not only in linear systems of algebraic equations and the geometry of  $n$ -dimensional Euclidean space, but also in the analysis of linear differential equations, linear boundary value problems, Fourier analysis, signal processing, numerical methods, and many, many other fields. Therefore, in order to master modern linear algebra and its applications, the first order of business is to acquire a firm understanding of fundamental vector space constructions.

One of the grand themes of mathematics is the recognition that many seemingly unrelated entities are, in fact, different manifestations of the same underlying abstract structure. This serves to unify and simplify the disparate special situations, at the expense of introducing an extra level of abstraction. Indeed, the history of mathematics, as well as your entire mathematical educational career, can be viewed as an evolution towards ever greater abstraction resulting in ever greater power for solving problems. Here, the abstract notion of a vector space serves to unify spaces of ordinary vectors, spaces of functions, such as polynomials, exponentials, and trigonometric functions, as well as spaces of matrices, spaces of linear operators, and so on, all in a common conceptual framework. Moreover, proofs that might appear to be complicated in a particular context often turn out to be relatively transparent when recast in the more inclusive vector space language. The price that one pays for the increased level of abstraction is that, while the underlying mathematics is not all that complicated, novices typically take a long time to assimilate the underlying concepts. In our opinion, the best way to approach the subject is to think in terms of concrete examples. First, make sure you understand what is being said in the case of ordinary Euclidean space. Once this is grasped, the next important case to consider is an elementary function space, e.g., the space of continuous scalar functions. With the two most important cases firmly in hand, the leap to the general abstract formulation should not be too painful. Patience is essential; ultimately, the only way to truly understand an abstract concept like a vector space is by working with it in real-life applications! And always keep in mind that the effort expended here will be amply rewarded later on.

Following an introduction to vector spaces and subspaces, we develop the fundamental notions of span and linear independence, first in the context of ordinary vectors, but then in more generality, with an emphasis on function spaces. These are then combined into the all-important definition of a basis of a vector space, leading to a linear algebraic characterization of its dimension. Here is where the distinction between finite-dimensional and infinite-dimensional vector spaces first becomes apparent, although the full ramifications of this dichotomy will take time to unfold. We will then study the four fundamental subspaces associated with a matrix — its image, kernel, coimage, and cokernel — and explain how they help us understand the structure and the solutions of linear algebraic systems. Of particular significance is the linear superposition principle that enables us to combine

solutions to linear systems. Superposition is the hallmark of linearity, and will apply not only to linear algebraic equations, but also to linear ordinary differential equations, linear boundary value problems, linear partial differential equations, linear integral equations, linear control systems, etc. The final section in this chapter develops some interesting applications in graph theory that serve to illustrate the fundamental matrix subspaces; these results will be developed further in our study of electrical circuits.

## 2.1 Real Vector Spaces

A vector space is the abstract reformulation of the quintessential properties of  $n$ -dimensional<sup>†</sup> *Euclidean space*  $\mathbb{R}^n$ , which is defined as the set of all real (column) vectors with  $n$  entries. The basic laws of vector addition and scalar multiplication in  $\mathbb{R}^n$  serve as the template for the following general definition.

**Definition 2.1.** A *vector space* is a set  $V$  equipped with two operations:

- (i) *Addition*: adding any pair of vectors  $\mathbf{v}, \mathbf{w} \in V$  produces another vector  $\mathbf{v} + \mathbf{w} \in V$ ;
- (ii) *Scalar Multiplication*: multiplying a vector  $\mathbf{v} \in V$  by a scalar  $c \in \mathbb{R}$  produces a vector  $c\mathbf{v} \in V$ .

These are subject to the following axioms, valid for all  $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$  and all scalars  $c, d \in \mathbb{R}$ :

- (a) *Commutativity of Addition*:  $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ .
- (b) *Associativity of Addition*:  $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$ .
- (c) *Additive Identity*: There is a zero element  $\mathbf{0} \in V$  satisfying  $\mathbf{v} + \mathbf{0} = \mathbf{v} = \mathbf{0} + \mathbf{v}$ .
- (d) *Additive Inverse*: For each  $\mathbf{v} \in V$  there is an element  $-\mathbf{v} \in V$  such that  $\mathbf{v} + (-\mathbf{v}) = \mathbf{0} = (-\mathbf{v}) + \mathbf{v}$ .
- (e) *Distributivity*:  $(c + d)\mathbf{v} = (c\mathbf{v}) + (d\mathbf{v})$ , and  $c(\mathbf{v} + \mathbf{w}) = (c\mathbf{v}) + (c\mathbf{w})$ .
- (f) *Associativity of Scalar Multiplication*:  $c(d\mathbf{v}) = (cd)\mathbf{v}$ .
- (g) *Unit for Scalar Multiplication*: the scalar  $1 \in \mathbb{R}$  satisfies  $1\mathbf{v} = \mathbf{v}$ .

**Remark.** For most of this text, we will deal with real vector spaces, in which the scalars are ordinary real numbers, as indicated in the definition. Complex vector spaces, where complex scalars are allowed, will be introduced in Section 3.6. Vector spaces over other fields are studied in abstract algebra, [38].

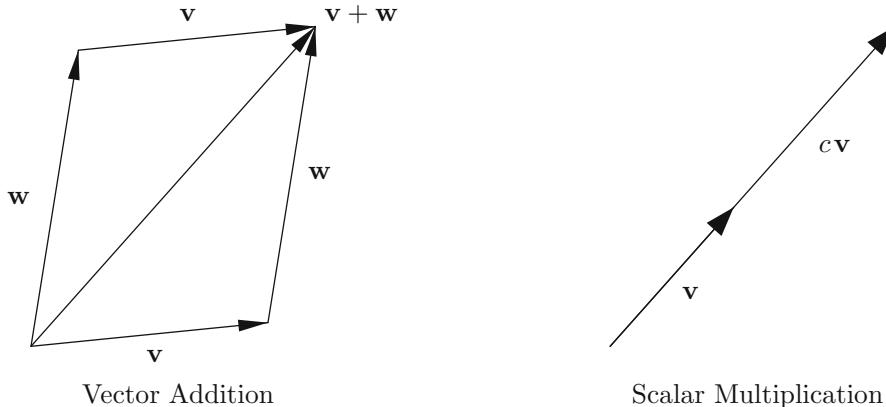
In the beginning, we will refer to the individual elements of a vector space as “vectors”, even though, as we shall see, they might also be functions, or matrices, or even more general objects. Unless we are dealing with certain specific examples such as a space of functions or matrices, we will use bold face, lower case Latin letters  $\mathbf{v}, \mathbf{w}, \dots$  to denote the elements of our vector space. We will usually use a bold face  $\mathbf{0}$  to denote the unique<sup>‡</sup> zero element of our vector space, while ordinary 0 denotes the real number zero.

The following identities are elementary consequences of the vector space axioms:

- (h)  $0\mathbf{v} = \mathbf{0}$ ;
- (i)  $(-1)\mathbf{v} = -\mathbf{v}$ ;
- (j)  $c\mathbf{0} = \mathbf{0}$ ;
- (k) If  $c\mathbf{v} = \mathbf{0}$ , then either  $c = 0$  or  $\mathbf{v} = \mathbf{0}$ .

<sup>†</sup> The precise definition of dimension will appear later, in Theorem 2.29.

<sup>‡</sup> See Exercise 2.1.12.

Figure 2.1. Vector Space Operations in  $\mathbb{R}^n$ .

Let us, for example, prove (h). Let  $\mathbf{z} = 0\mathbf{v}$ . Then, by the distributive property,

$$\mathbf{z} + \mathbf{z} = 0\mathbf{v} + 0\mathbf{v} = (0 + 0)\mathbf{v} = 0\mathbf{v} = \mathbf{z}.$$

Adding  $-\mathbf{z}$  to both sides of this equation, and making use of axioms (b), (d), and then (c), we conclude that

$$\mathbf{z} = \mathbf{z} + \mathbf{0} = \mathbf{z} + (\mathbf{z} + (-\mathbf{z})) = (\mathbf{z} + \mathbf{z}) + (-\mathbf{z}) = \mathbf{z} + (-\mathbf{z}) = \mathbf{0},$$

which completes the proof. Verification of the other three properties is left as an exercise for the reader.

Let us now introduce the most important examples of (real) vector spaces.

**Example 2.2.** As noted above, the prototypical example of a real vector space is the Euclidean space  $\mathbb{R}^n$ , consisting of all  $n$ -tuples of real numbers  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$ , which we consistently write as column vectors. Vector addition and scalar multiplication are defined in the usual manner:

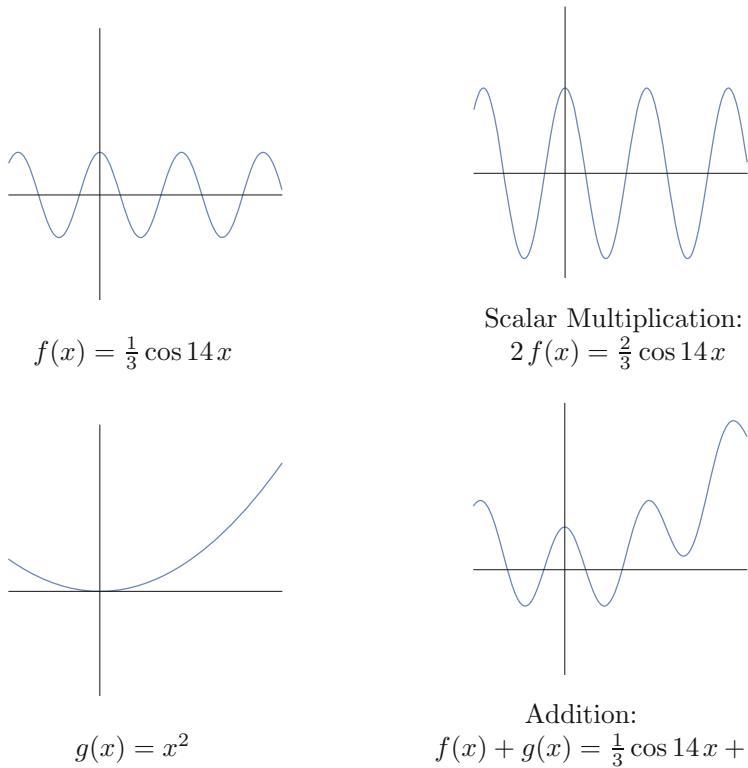
$$\mathbf{v} + \mathbf{w} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{pmatrix}, \quad c\mathbf{v} = \begin{pmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_n \end{pmatrix}, \quad \text{whenever } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}, \quad c \in \mathbb{R}.$$

The zero vector is  $\mathbf{0} = (0, \dots, 0)^T$ . The two vector space operations are illustrated in Figure 2.1. The fact that vectors in  $\mathbb{R}^n$  satisfy all of the vector space axioms is an immediate consequence of the laws of vector addition and scalar multiplication.

**Example 2.3.** Let  $\mathcal{M}_{m \times n}$  denote the space of all real matrices of size  $m \times n$ . Then  $\mathcal{M}_{m \times n}$  forms a vector space under the laws of matrix addition and scalar multiplication. The zero element is the zero matrix  $\mathbf{0}$ . (We are ignoring matrix multiplication, which is *not* a vector space operation.) Again, the vector space axioms are immediate consequences of the basic laws of matrix arithmetic. The preceding example of the vector space  $\mathbb{R}^n = \mathcal{M}_{n \times 1}$  is a particular case in which the matrices have only one column.

**Example 2.4.** Consider the space

$$\mathcal{P}^{(n)} = \{ p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 \} \quad (2.1)$$



**Figure 2.2.** Vector Space Operations in Function Space.

consisting of all real polynomials of degree  $\leq n$ . Addition of polynomials is defined in the usual manner; for example,

$$(x^2 - 3x) + (2x^2 - 5x + 4) = 3x^2 - 8x + 4.$$

Note that the sum  $p(x) + q(x)$  of two polynomials of degree  $\leq n$  also has degree  $\leq n$ . The zero element of  $\mathcal{P}^{(n)}$  is the zero polynomial. We can multiply polynomials by scalars — real constants — in the usual fashion; for example if  $p(x) = x^2 - 2x$ , then  $3p(x) = 3x^2 - 6x$ . The proof that  $\mathcal{P}^{(n)}$  satisfies the vector space axioms is an easy consequence of the basic laws of polynomial algebra.

**Warning.** It is not true that the sum of two polynomials of degree  $n$  also has degree  $n$ . For example  $(x^2 + 1) + (-x^2 + x) = x + 1$  has degree 1 even though the two summands have degree 2. This means that the set of polynomials of degree  $= n$  is *not* a vector space.

**Warning.** You might be tempted to identify a scalar with a constant polynomial, but one should really regard these as two completely different objects — one is a *number*, while the other is a *constant function*. To add to the confusion, one typically uses the same notation for these two objects; for instance, 0 could mean either the real number 0 or the constant function taking the value 0 everywhere, which is the zero element, **0**, of this vector space. The reader needs to exercise due care when interpreting each occurrence.

For much of analysis, including differential equations, Fourier theory, numerical methods, etc., the most important vector spaces consist of functions that have certain prescribed properties. The simplest such example is the following.

**Example 2.5.** Let  $I \subset \mathbb{R}$  be an interval<sup>†</sup>. Consider the *function space*  $\mathcal{F} = \mathcal{F}(I)$  whose elements are all real-valued functions  $f(x)$  defined for all  $x \in I$ . The claim is that the function space  $\mathcal{F}$  has the structure of a vector space. Addition of functions in  $\mathcal{F}$  is defined in the usual manner:  $(f+g)(x) = f(x) + g(x)$  for all  $x \in I$ . Multiplication by scalars  $c \in \mathbb{R}$  is the same as multiplication by constants,  $(c f)(x) = c f(x)$ . The zero element is the zero function — the constant function that is identically 0 for all  $x \in I$ . The proof of the vector space axioms is straightforward. Observe that we are ignoring all additional operations that affect functions such as multiplication, division, inversion, composition, etc.; these are irrelevant as far as the vector space structure of  $\mathcal{F}$  goes.

**Example 2.6.** The preceding examples are all, in fact, special cases of an even more general construction. A clue is to note that the last example of a function space does not make any use of the fact that the domain of the functions is a real interval. Indeed, the same construction produces a function space  $\mathcal{F}(I)$  corresponding to *any* subset  $I \subset \mathbb{R}$ .

Even more generally, let  $S$  be *any* set. Let  $\mathcal{F} = \mathcal{F}(S)$  denote the space of all real-valued functions  $f: S \rightarrow \mathbb{R}$ . Then we claim that  $V$  is a vector space under the operations of function addition and scalar multiplication. More precisely, given functions  $f$  and  $g$ , we define their sum to be the function  $h = f + g$  such that  $h(x) = f(x) + g(x)$  for all  $x \in S$ . Similarly, given a function  $f$  and a real scalar  $c \in \mathbb{R}$ , we define the scalar multiple  $g = c f$  to be the function such that  $g(x) = c f(x)$  for all  $x \in S$ . The proof of the vector space axioms is straightforward, and the reader should be able to fill in the necessary details.

In particular, if  $S \subset \mathbb{R}$  is an interval, then  $\mathcal{F}(S)$  coincides with the space of scalar functions described in the preceding example. If  $S \subset \mathbb{R}^n$  is a subset of Euclidean space, then the elements of  $\mathcal{F}(S)$  are real-valued functions  $f(x_1, \dots, x_n)$  depending upon the  $n$  variables corresponding to the coordinates of points  $\mathbf{x} = (x_1, \dots, x_n) \in S$  in the domain. In this fashion, the set of real-valued functions defined on any domain in  $\mathbb{R}^n$  forms a vector space.

Another useful example is to let  $S = \{x_1, \dots, x_n\} \subset \mathbb{R}$  be a finite set of real numbers. A real-valued function  $f: S \rightarrow \mathbb{R}$  is defined by its values  $f(x_1), f(x_2), \dots, f(x_n)$  at the specified points. In applications, these objects serve to indicate the *sample values* of a scalar function  $f(x) \in \mathcal{F}(\mathbb{R})$  taken at the *sample points*  $x_1, \dots, x_n$ . For example, if  $f(x) = x^2$  and the sample points are  $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3$ , then the corresponding sample values are  $f(x_1) = 0, f(x_2) = 1, f(x_3) = 4, f(x_4) = 9$ . When measuring a physical quantity — velocity, temperature, pressure, etc. — one typically records only a finite set of sample values. The intermediate, non-recorded values between the sample points are then reconstructed through some form of interpolation, a topic that we shall visit in depth in Chapters 4 and 5.

Interestingly, the sample values  $f(x_i)$  can be identified with the entries  $f_i$  of a vector

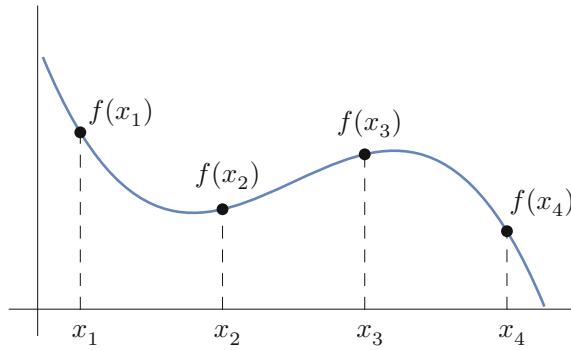
$$\mathbf{f} = (f_1, f_2, \dots, f_n)^T = (f(x_1), f(x_2), \dots, f(x_n))^T \in \mathbb{R}^n,$$

---

<sup>†</sup> An *interval* is a subset  $I \subset \mathbb{R}$  that contains all the real numbers between  $a, b \in \mathbb{R}$ , where  $a < b$ , and can be

- *closed*, meaning that it includes its endpoints:  $I = [a, b] = \{x \mid a \leq x \leq b\}$ ;
- *open*, which does not include either endpoint:  $I = (a, b) = \{x \mid a < x < b\}$ ; or
- *half-open*, which includes one but not the other endpoint, so  $I = [a, b) = \{x \mid a \leq x < b\}$  or  $I = (a, b] = \{x \mid a < x \leq b\}$ .

An open endpoint is allowed to be infinite; in particular,  $(-\infty, \infty) = \mathbb{R}$  is another way of writing the entire real line.



**Figure 2.3.** Sampled Function.

known as the *sample vector*. Every sampled function  $f: S \rightarrow \mathbb{R}$  corresponds to a unique vector  $\mathbf{f} \in \mathbb{R}^n$  and vice versa. (But keep in mind that different scalar functions  $f(x) \in \mathcal{F}(\mathbb{R})$  may have the same sample values.) Addition of sample functions corresponds to addition of their sample vectors, as does scalar multiplication. Thus, *the vector space of sample functions  $\mathcal{F}(S) = \mathcal{F}\{x_1, \dots, x_n\}$  is the same as the vector space  $\mathbb{R}^n$ !* The identification of sampled functions as vectors is of fundamental importance in modern signal processing and data analysis, as we will see below.

**Example 2.7.** The above construction admits yet a further generalization. We continue to let  $S$  be an arbitrary set. Let  $V$  be a vector space. The claim is that the space  $\mathcal{F}(S, V)$  consisting of all  $V$ -valued functions  $\mathbf{f}: S \rightarrow V$  is a vector space. In other words, we replace the particular vector space  $\mathbb{R}$  in the preceding example by a general vector space  $V$ , and the same conclusion holds. The operations of function addition and scalar multiplication are defined in the evident manner:  $(\mathbf{f} + \mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x)$  and  $(c\mathbf{f})(x) = c\mathbf{f}(x)$  for  $x \in S$ , where we are using the vector addition and scalar multiplication operations on  $V$  to induce corresponding operations on  $V$ -valued functions. The proof that  $\mathcal{F}(S, V)$  satisfies all of the vector space axioms proceeds as before.

The most important example of such a function space arises when  $S \subset \mathbb{R}^n$  is a domain in Euclidean space and  $V = \mathbb{R}^m$  is itself a Euclidean space. In this case, the elements of  $\mathcal{F}(S, \mathbb{R}^m)$  consist of vector-valued functions  $\mathbf{f}: S \rightarrow \mathbb{R}^m$ , so that  $\mathbf{f}(\mathbf{x}) = (f_1(x_1, \dots, x_n), \dots, f_m(x_1, \dots, x_n))^T$  is a column vector consisting of  $m$  functions of  $n$  variables, all defined on a common domain  $S$ . The general construction implies that addition and scalar multiplication of vector-valued functions is done componentwise; for example

$$2 \begin{pmatrix} x^2 \\ e^x - 4 \end{pmatrix} - \begin{pmatrix} \cos x \\ x \end{pmatrix} = \begin{pmatrix} 2x^2 - \cos x \\ 2e^x - x - 8 \end{pmatrix}.$$

Of particular importance are the vector fields arising in physics, including gravitational force fields, electromagnetic fields, fluid velocity fields, and many others.

## Exercises

- 2.1.1. Show that the set of complex numbers  $x + iy$  forms a real vector space under the operations of addition  $(x + iy) + (u + iv) = (x + u) + i(y + v)$  and scalar multiplication  $c(x + iy) = cx + icy$ . (But complex multiplication is *not* a real vector space operation.)

2.1.2. Show that the positive quadrant  $Q = \{(x, y) \mid x, y > 0\} \subset \mathbb{R}^2$  forms a vector space if we define addition by  $(x_1, y_1) + (x_2, y_2) = (x_1 x_2, y_1 y_2)$  and scalar multiplication by  $c(x, y) = (x^c, y^c)$ .

◇ 2.1.3. Let  $S$  be any set. Carefully justify the validity of all the vector space axioms for the space  $\mathcal{F}(S)$  consisting of all real-valued functions  $f: S \rightarrow \mathbb{R}$ .

2.1.4. Let  $S = \{0, 1, 2, 3\}$ . (a) Find the sample vectors corresponding to the functions  $1$ ,  $\cos \pi x$ ,  $\cos 2\pi x$ ,  $\cos 3\pi x$ . (b) Is a function uniquely determined by its sample values?

2.1.5. Find two different functions  $f(x)$  and  $g(x)$  that have the *same* sample vectors  $\mathbf{f}, \mathbf{g}$  at the sample points  $x_1 = 0$ ,  $x_2 = 1$ ,  $x_3 = -1$ .

2.1.6. (a) Let  $x_1 = 0$ ,  $x_2 = 1$ . Find the unique linear function  $f(x) = ax + b$  that has the sample vector  $\mathbf{f} = (3, -1)^T$ . (b) Let  $x_1 = 0$ ,  $x_2 = 1$ ,  $x_3 = -1$ . Find the unique quadratic function  $f(x) = ax^2 + bx + c$  with sample vector  $\mathbf{f} = (1, -2, 0)^T$ .

2.1.7. Let  $\mathcal{F}(\mathbb{R}^2, \mathbb{R}^2)$  denote the vector space consisting of all functions  $\mathbf{f}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ .

(a) Which of the following functions  $\mathbf{f}(x, y)$  are elements? (i)  $x^2 + y^2$ , (ii)  $\begin{pmatrix} x - y \\ xy \end{pmatrix}$ , (iii)  $\begin{pmatrix} e^x \\ \cos y \end{pmatrix}$ , (iv)  $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ , (v)  $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$ , (vi)  $\begin{pmatrix} x \\ y \\ x + y \end{pmatrix}$ . (b) Sum all of the elements of  $\mathcal{F}(\mathbb{R}^2, \mathbb{R}^2)$  you identified in part (a). Then multiply your sum by the scalar  $-5$ .  
(c) Carefully describe the zero element of the vector space  $\mathcal{F}(\mathbb{R}^2, \mathbb{R}^2)$ .

◇ 2.1.8. A *planar vector field* is a function that assigns a vector  $\mathbf{v}(x, y) = \begin{pmatrix} v_1(x, y) \\ v_2(x, y) \end{pmatrix}$  to each point  $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2$ . Explain why the set of all planar vector fields forms a vector space.

◇ 2.1.9. Let  $h, k > 0$  be fixed. Let  $S = \{(ih, jk) \mid 1 \leq i \leq m, 1 \leq j \leq n\}$  be points in a rectangular planar grid. Show that the function space  $\mathcal{F}(S)$  can be identified with the vector space of  $m \times n$  matrices  $\mathcal{M}_{m \times n}$ .

2.1.10. The space  $\mathbb{R}^\infty$  is defined as the set of all infinite real sequences  $\mathbf{a} = (a_1, a_2, a_3, \dots)$ , where  $a_i \in \mathbb{R}$ . Define addition and scalar multiplication in such a way as to make  $\mathbb{R}^\infty$  into a vector space. Explain why all the vector space axioms are valid.

2.1.11. Prove the basic vector space properties (i), (j), (k) following Definition 2.1.

◇ 2.1.12. Prove that a vector space has only one zero element  $\mathbf{0}$ .

◇ 2.1.13. Suppose that  $V$  and  $W$  are vector spaces. The *Cartesian product space*, denoted by  $V \times W$ , is defined as the set of all ordered pairs  $(\mathbf{v}, \mathbf{w})$ , where  $\mathbf{v} \in V, \mathbf{w} \in W$ , with vector addition  $(\mathbf{v}, \mathbf{w}) + (\hat{\mathbf{v}}, \hat{\mathbf{w}}) = (\mathbf{v} + \hat{\mathbf{v}}, \mathbf{w} + \hat{\mathbf{w}})$  and scalar multiplication  $c(\mathbf{v}, \mathbf{w}) = (c\mathbf{v}, c\mathbf{w})$ .  
(a) Prove that  $V \times W$  is a vector space. (b) Explain why  $\mathbb{R} \times \mathbb{R}$  is the same as  $\mathbb{R}^2$ .  
(c) More generally, explain why  $\mathbb{R}^m \times \mathbb{R}^n$  is the same as  $\mathbb{R}^{m+n}$ .

2.1.14. Use Exercise 2.1.13 to show that the space of pairs  $(f(x), a)$ , where  $f$  is a continuous scalar function and  $a$  is a real number, is a vector space. What is the zero element? Be precise! Write out the laws of vector addition and scalar multiplication.

## 2.2 Subspaces

In the preceding section, we were introduced to the most basic vector spaces that arise in this text. Almost all of the vector spaces used in applications appear as subsets of these prototypical examples.

**Definition 2.8.** A *subspace* of a vector space  $V$  is a subset  $W \subset V$  that is a vector space in its own right — under the same operations of vector addition and scalar multiplication and the same zero element.

In particular, a subspace  $W$  *must* contain the zero element of  $V$ . Proving that a given subset of a vector space is a subspace is particularly easy: we need only check its *closure* under addition and scalar multiplication.

**Proposition 2.9.** A nonempty subset  $W \subset V$  of a vector space is a subspace if and only if

- (a) for every  $\mathbf{v}, \mathbf{w} \in W$ , the sum  $\mathbf{v} + \mathbf{w} \in W$ , and
- (b) for every  $\mathbf{v} \in W$  and every  $c \in \mathbb{R}$ , the scalar product  $c\mathbf{v} \in W$ .

*Proof:* The proof is immediate. For example, let us check commutativity. The subspace elements  $\mathbf{v}, \mathbf{w} \in W$  can be regarded as elements of  $V$ , in which case  $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$  because  $V$  is a vector space. But the closure condition implies that the common sum also belongs to  $W$ , and so the commutativity axiom also holds for elements of  $W$ . Establishing the validity of the other axioms is equally easy. *Q.E.D.*

It is sometimes convenient to combine the two closure conditions. Thus, to prove that  $W \subset V$  is a subspace, it suffices to check that  $c\mathbf{v} + d\mathbf{w} \in W$  for all  $\mathbf{v}, \mathbf{w} \in W$  and  $c, d \in \mathbb{R}$ .

**Example 2.10.** Let us list some examples of subspaces of the three-dimensional Euclidean space  $\mathbb{R}^3$ .

- (a) The trivial subspace  $W = \{\mathbf{0}\}$ . Demonstrating closure is easy: since there is only one element  $\mathbf{0}$  in  $W$ , we just need to check that  $\mathbf{0} + \mathbf{0} = \mathbf{0} \in W$  and  $c\mathbf{0} = \mathbf{0} \in W$  for every scalar  $c$ .
- (b) The entire space  $W = \mathbb{R}^3$ . Here closure is immediate because  $\mathbb{R}^3$  is a vector space in its own right.
- (c) The set of all vectors of the form  $(x, y, 0)^T$ , i.e., the  $xy$  coordinate plane. To prove closure, we check that all sums  $(x, y, 0)^T + (\hat{x}, \hat{y}, 0)^T = (x + \hat{x}, y + \hat{y}, 0)^T$  and scalar multiples  $c(x, y, 0)^T = (cx, cy, 0)^T$  of vectors in the  $xy$ -plane remain in the plane.
- (d) The set of solutions  $(x, y, z)^T$  to the homogeneous linear equation

$$3x + 2y - z = 0. \quad (2.2)$$

Indeed, if  $\mathbf{x} = (x, y, z)^T$  is a solution, then so is every scalar multiple  $c\mathbf{x} = (cx, cy, cz)^T$  since

$$3(cx) + 2(cy) - (cz) = c(3x + 2y - z) = 0.$$

Moreover, if  $\hat{\mathbf{x}} = (\hat{x}, \hat{y}, \hat{z})^T$  is a second solution, the sum  $\mathbf{x} + \hat{\mathbf{x}} = (x + \hat{x}, y + \hat{y}, z + \hat{z})^T$  is also a solution, since

$$3(x + \hat{x}) + 2(y + \hat{y}) - (z + \hat{z}) = (3x + 2y - z) + (3\hat{x} + 2\hat{y} - \hat{z}) = 0.$$

The solution space is, in fact, the two-dimensional plane passing through the origin with normal vector  $(3, 2, -1)^T$ .

- (e) The set of all vectors lying in the plane spanned by the vectors  $\mathbf{v}_1 = (2, -3, 0)^T$  and  $\mathbf{v}_2 = (1, 0, 3)^T$ . In other words, we consider all vectors of the form

$$\mathbf{v} = a\mathbf{v}_1 + b\mathbf{v}_2 = a \begin{pmatrix} 2 \\ -3 \\ 0 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 2a + b \\ -3a \\ 3b \end{pmatrix},$$

where  $a, b \in \mathbb{R}$  are arbitrary scalars. If  $\mathbf{v} = a\mathbf{v}_1 + b\mathbf{v}_2$  and  $\mathbf{w} = \hat{a}\mathbf{v}_1 + \hat{b}\mathbf{v}_2$  are any two vectors in the span, then so is

$$c\mathbf{v} + d\mathbf{w} = c(a\mathbf{v}_1 + b\mathbf{v}_2) + d(\hat{a}\mathbf{v}_1 + \hat{b}\mathbf{v}_2) = (ac + \hat{a}d)\mathbf{v}_1 + (bc + \hat{b}d)\mathbf{v}_2 = \tilde{a}\mathbf{v}_1 + \tilde{b}\mathbf{v}_2,$$

where  $\tilde{a} = ac + \hat{a}d$ ,  $\tilde{b} = bc + \hat{b}d$ . This demonstrates that the span is a subspace of  $\mathbb{R}^3$ . The reader might already have noticed that this subspace is the same plane defined by (2.2).

**Example 2.11.** The following subsets of  $\mathbb{R}^3$  are *not* subspaces.

- (a) The set  $P$  of all vectors of the form  $(x, y, 1)^T$ , i.e., the plane parallel to the  $xy$  coordinate plane passing through  $(0, 0, 1)^T$ . Indeed,  $(0, 0, 0)^T \notin P$ , which is the most basic requirement for a subspace. In fact, neither of the closure axioms hold for this subset.
- (b) The nonnegative orthant  $\mathcal{O}^+ = \{x \geq 0, y \geq 0, z \geq 0\}$ . Although  $\mathbf{0} \in \mathcal{O}^+$ , and the sum of two vectors in  $\mathcal{O}^+$  also belongs to  $\mathcal{O}^+$ , multiplying by negative scalars takes us outside the orthant, violating closure under scalar multiplication.
- (c) The unit sphere  $S_1 = \{x^2 + y^2 + z^2 = 1\}$ . Again,  $\mathbf{0} \notin S_1$ . More generally, curved surfaces, such as the paraboloid  $P = \{z = x^2 + y^2\}$ , are not subspaces. Although  $\mathbf{0} \in P$ , most scalar multiples of elements of  $P$  do not belong to  $P$ . For example,  $(1, 1, 2)^T \in P$ , but  $2(1, 1, 2)^T = (2, 2, 4)^T \notin P$ .

In fact, there are only four fundamentally different types of subspaces  $W \subset \mathbb{R}^3$  of three-dimensional Euclidean space:

- (i) the entire three-dimensional space  $W = \mathbb{R}^3$ ,
- (ii) a plane passing through the origin,
- (iii) a line passing through the origin,
- (iv) a point — the trivial subspace  $W = \{\mathbf{0}\}$ .

We can establish this observation by the following argument. If  $W = \{\mathbf{0}\}$  contains only the zero vector, then we are in case (iv). Otherwise,  $W \subset \mathbb{R}^3$  contains a nonzero vector  $\mathbf{0} \neq \mathbf{v}_1 \in W$ . But since  $W$  must contain all scalar multiples  $c\mathbf{v}_1$  of this element, it includes the entire line in the direction of  $\mathbf{v}_1$ . If  $W$  contains another vector  $\mathbf{v}_2$  that does not lie in the line through  $\mathbf{v}_1$ , then it must contain the entire plane  $\{c\mathbf{v}_1 + d\mathbf{v}_2\}$  spanned by  $\mathbf{v}_1, \mathbf{v}_2$ . Finally, if there is a third vector  $\mathbf{v}_3$  not contained in this plane, then we claim that  $W = \mathbb{R}^3$ . This final fact will be an immediate consequence of general results in this chapter, although the interested reader might try to prove it directly before proceeding.

**Example 2.12.** Let  $I \subset \mathbb{R}$  be an interval, and let  $\mathcal{F}(I)$  be the space of real-valued functions  $f: I \rightarrow \mathbb{R}$ . Let us look at some of the most important examples of subspaces of  $\mathcal{F}(I)$ . In each case, we need only verify the closure conditions to verify that the given subset is indeed a subspace. In particular, the zero function belongs to each of the subspaces.

- (a) The space  $\mathcal{P}^{(n)}$  of polynomials of degree  $\leq n$ , which we already encountered.
- (b) The space  $\mathcal{P}^{(\infty)} = \bigcup_{n \geq 0} \mathcal{P}^{(n)}$  consisting of all polynomials. Closure means that the sum of any two polynomials is a polynomial, as is any scalar (constant) multiple of a polynomial.
- (c) The space  $C^0(I)$  of all continuous functions. Closure of this subspace relies on knowing that if  $f(x)$  and  $g(x)$  are continuous, then both  $f(x) + g(x)$  and  $cf(x)$ , for any  $c \in \mathbb{R}$ , are also continuous — two basic results from calculus, [2, 78].

(d) More restrictively, we can consider the subspace  $C^n(I)$  consisting of all functions  $f(x)$  that have  $n$  continuous derivatives  $f'(x), f''(x), \dots, f^{(n)}(x)$  on<sup>†</sup>  $I$ . Again, we need to know that if  $f(x)$  and  $g(x)$  have  $n$  continuous derivatives, then so does  $cf(x) + dg(x)$  for all  $c, d \in \mathbb{R}$ .

(e) The space  $C^\infty(I) = \bigcap_{n \geq 0} C^n(I)$  of infinitely differentiable or *smooth* functions is also a subspace. This can be proved directly, or it follows from the general fact that the intersection of subspaces is a subspace, cf. Exercise 2.2.23.

(f) The space  $\mathcal{A}(I)$  of analytic functions on the interval  $I$ . Recall that a function  $f(x)$  is called *analytic* at a point  $a$  if it is smooth, and, moreover, its Taylor series

$$f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \dots = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n \quad (2.3)$$

converges to  $f(x)$  for all  $x$  sufficiently close to  $a$ . (The series is not required to converge on the entire interval  $I$ .) Not every smooth function is analytic, and so  $\mathcal{A}(I) \subsetneq C^\infty(I)$ . An explicit example of a smooth but non-analytic function can be found in Exercise 2.2.30.

(g) The set of all mean zero functions. The *mean* or *average* of an integrable function defined on a closed interval  $I = [a, b]$  is the real number

$$\bar{f} = \frac{1}{b-a} \int_a^b f(x) dx. \quad (2.4)$$

In particular,  $f$  has *mean zero* if and only if  $\int_a^b f(x) dx = 0$ . Since  $\overline{f+g} = \bar{f} + \bar{g}$ , and  $\overline{cf} = c\bar{f}$ , sums and scalar multiples of mean zero functions also have mean zero, proving closure.

(h) Fix  $x_0 \in I$ . The set of all functions  $f(x)$  that vanish at the point,  $f(x_0) = 0$ , is a subspace. Indeed, if  $f(x_0) = 0$  and  $g(x_0) = 0$ , then, clearly  $(cf + dg)(x_0) = cf(x_0) + dg(x_0) = 0$  for all  $c, d \in \mathbb{R}$ , proving closure. This example can evidently be generalized to functions that vanish at several points, or even on an entire subset  $S \subset I$ .

(i) The set of all solutions  $u = f(x)$  to the homogeneous linear differential equation

$$u'' + 2u' - 3u = 0.$$

Indeed, if  $f(x)$  and  $g(x)$  are solutions, then so is  $f(x) + g(x)$  and  $cf(x)$  for all  $c \in \mathbb{R}$ . Note that we do *not* need to actually solve the equation to verify these claims! They follow directly from linearity:

$$\begin{aligned} (f+g)'' + 2(f+g)' - 3(f+g) &= (f'' + 2f' - 3f) + (g'' + 2g' - 3g) = 0, \\ (cf)'' + 2(cf)' - 3(cf) &= c(f'' + 2f' - 3f) = 0. \end{aligned}$$

**Remark.** In the last three examples, 0 is essential for the indicated set of functions to be a subspace. The set of functions such that  $f(x_0) = 1$ , say, is not a subspace. The set of functions with a given nonzero mean, say  $\bar{f} = 3$ , is also not a subspace. Nor is the set of solutions to an inhomogeneous ordinary differential equation, say  $u'' + 2u' - 3u = x - 3$ . None of these subsets contain the zero function, nor do they satisfy the closure conditions.

---

<sup>†</sup> We use one-sided derivatives at any endpoint belonging to the interval.

## Exercises

- 2.2.1. (a) Prove that the set of all vectors  $(x, y, z)^T$  such that  $x - y + 4z = 0$  forms a subspace of  $\mathbb{R}^3$ . (b) Explain why the set of all vectors that satisfy  $x - y + 4z = 1$  does not form a subspace.
- 2.2.2. Which of the following are subspaces of  $\mathbb{R}^3$ ? Justify your answers! (a) The set of all vectors  $(x, y, z)^T$  satisfying  $x + y + z + 1 = 0$ . (b) The set of vectors of the form  $(t, -t, 0)^T$  for  $t \in \mathbb{R}$ . (c) The set of vectors of the form  $(r - s, r + 2s, -s)^T$  for  $r, s \in \mathbb{R}$ . (d) The set of vectors whose first component equals 0. (e) The set of vectors whose last component equals 1. (f) The set of all vectors  $(x, y, z)^T$  with  $x \geq y \geq z$ . (g) The set of all solutions to the equation  $z = x - y$ . (h) The set of all solutions to  $z = xy$ . (i) The set of all solutions to  $x^2 + y^2 + z^2 = 0$ . (j) The set of all solutions to the system  $xy = yz = xz$ .
- 2.2.3. Graph the following subsets of  $\mathbb{R}^3$  and use this to explain which are subspaces:  
 (a) The line  $(t, -t, 3t)^T$  for  $t \in \mathbb{R}$ . (b) The helix  $(\cos t, \sin t, t)^T$ . (c) The surface  $x - 2y + 3z = 0$ . (d) The unit ball  $x^2 + y^2 + z^2 < 1$ . (e) The cylinder  $(y+2)^2 + (z-1)^2 = 5$ .  
 (f) The intersection of the cylinders  $(x-1)^2 + y^2 = 1$  and  $(x+1)^2 + y^2 = 1$ .
- 2.2.4. Show that if  $W \subset \mathbb{R}^3$  is a subspace containing the vectors  $(1, 2, -1)^T$ ,  $(2, 0, 1)^T$ ,  $(0, -1, 3)^T$ , then  $W = \mathbb{R}^3$ .
- 2.2.5. *True or false:* An interval is a vector space.
- 2.2.6. (a) Can you construct an example of a subset  $S \subset \mathbb{R}^2$  with the property that  $c\mathbf{v} \in S$  for all  $c \in \mathbb{R}$ ,  $\mathbf{v} \in S$ , and yet  $S$  is not a subspace? (b) What about an example in which  $\mathbf{v} + \mathbf{w} \in S$  for every  $\mathbf{v}, \mathbf{w} \in S$ , and yet  $S$  is not a subspace?
- 2.2.7. Determine which of the following sets of vectors  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  are subspaces of  $\mathbb{R}^n$ : (a) all equal entries  $x_1 = \dots = x_n$ ; (b) all positive entries:  $x_i \geq 0$ ; (c) first and last entries equal to zero:  $x_1 = x_n = 0$ ; (d) entries add up to zero:  $x_1 + \dots + x_n = 0$ ; (e) first and last entries differ by one:  $x_1 - x_n = 1$ .
- 2.2.8. Prove that the set of all solutions  $\mathbf{x}$  of the linear system  $A\mathbf{x} = \mathbf{b}$  forms a subspace if and only if the system is homogeneous.
- 2.2.9. A square matrix is called *strictly lower triangular* if all entries on or above the main diagonal are 0. Prove that the space of strictly lower triangular matrices is a subspace of the vector space of all  $n \times n$  matrices.
- 2.2.10. Which of the following are subspaces of the vector space of  $n \times n$  matrices  $\mathcal{M}_{n \times n}$ ?  
 The set of all (a) regular matrices; (b) nonsingular matrices; (c) singular matrices;  
 (d) lower triangular matrices; (e) lower unitriangular matrices; (f) diagonal matrices;  
 (g) symmetric matrices; (h) skew-symmetric matrices.
- ◇ 2.2.11. The *trace* of an  $n \times n$  matrix  $A \in \mathcal{M}_{n \times n}$  is defined to be the sum of its diagonal entries:  $\text{tr } A = a_{11} + a_{22} + \dots + a_{nn}$ . Prove that the set of trace zero matrices,  $\text{tr } A = 0$ , is a subspace of  $\mathcal{M}_{n \times n}$ .
- 2.2.12. (a) Is the set of  $n \times n$  matrices with  $\det A = 1$  a subspace of  $\mathcal{M}_{n \times n}$ ?  
 (b) What about the matrices with  $\det A = 0$ ?
- 2.2.13. Let  $V = C^0(\mathbb{R})$  be the vector space consisting of all continuous functions  $f: \mathbb{R} \rightarrow \mathbb{R}$ . Explain why the set of all functions such that  $f(1) = 0$  is a subspace, but the set of functions such that  $f(0) = 1$  is not. For which values of  $a, b$  does the set of functions such that  $f(a) = b$  form a subspace?

- 2.2.14. Which of the following are vector spaces? Justify your answer! (a) The set of all row vectors of the form  $(a, 3a)$ . (b) The set of all vectors of the form  $(a, a+1)$ . (c) The set of all continuous functions for which  $f(-1) = 0$ . (d) The set of all periodic functions of period 1, i.e.,  $f(x+1) = f(x)$ . (e) The set of all non-negative functions:  $f(x) \geq 0$ . (f) The set of all even polynomials:  $p(x) = p(-x)$ . (g) The set of all polynomials  $p(x)$  that have  $x-1$  as a factor. (h) The set of all quadratic forms  $q(x, y) = ax^2 + bxy + cy^2$ .
- 2.2.15. Determine which of the following conditions describe subspaces of the vector space  $C^1$  consisting of all continuously differentiable scalar functions  $f(x)$ .  
 (a)  $f(2) = f(3)$ , (b)  $f'(2) = f(3)$ , (c)  $f'(x) + f(x) = 0$ , (d)  $f(2-x) = f(x)$ ,  
 (e)  $f(x+2) = f(x) + 2$ , (f)  $f(-x) = e^x f(x)$ . (g)  $f(x) = a + b|x|$  for some  $a, b \in \mathbb{R}$ ,
- 2.2.16. Let  $V = C^0[a, b]$  be the vector space consisting of all functions  $f(t)$  that are defined and continuous on the interval  $0 \leq t \leq 1$ . Which of the following conditions define subspaces of  $V$ ? Explain your answer. (a)  $f(0) = 0$ , (b)  $f(0) = 2f(1)$ , (c)  $f(0)f(1) = 1$ , (d)  $f(0) = 0$  or  $f(1) = 0$ , (e)  $f(1-t) = -t f(t)$ , (f)  $f(1-t) = 1 - f(t)$ ,  
 (g)  $f\left(\frac{1}{2}\right) = \int_0^1 f(t) dt$ , (h)  $\int_0^1 (t-1) f(t) dt = 0$ , (i)  $\int_0^t f(s) \sin s ds = \sin t$ .
- 2.2.17. Prove that the set of solutions to the second order ordinary differential equation  $u'' = xu$  is a vector space.
- 2.2.18. Show that the set of solutions to  $u'' = x + u$  does not form a vector space.
- 2.2.19. (a) Prove that  $C^1([a, b], \mathbb{R}^2)$ , which is the space of continuously differentiable parameterized plane curves  $\mathbf{f}: [a, b] \rightarrow \mathbb{R}^2$ , is a vector space.  
 (b) Is the subset consisting of all curves that go through the origin a subspace?
- 2.2.20. A *planar vector field*  $\mathbf{v}(x, y) = (u(x, y), v(x, y))^T$  is called *irrotational* if it has zero divergence:  $\nabla \cdot \mathbf{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \equiv 0$ . Prove that the set of all irrotational vector fields is a subspace of the space of all planar vector fields.
- 2.2.21. Let  $C \subset \mathbb{R}^\infty$  denote the set of all convergent sequences of real numbers, where  $\mathbb{R}^\infty$  was defined in Exercise 2.2.21. Is  $C$  a subspace?
- ◊ 2.2.22. Show that if  $W$  and  $Z$  are subspaces of  $V$ , then (a) their intersection  $W \cap Z$  is a subspace of  $V$ , (b) their sum  $W + Z = \{\mathbf{w} + \mathbf{z} \mid \mathbf{w} \in W, \mathbf{z} \in Z\}$  is also a subspace, but (c) their union  $W \cup Z$  is not a subspace of  $V$ , unless  $W \subset Z$  or  $Z \subset W$ .
- ◊ 2.2.23. Let  $V$  be a vector space. Prove that the intersection  $\bigcap W_i$  of any collection (finite or infinite) of subspaces  $W_i \subset V$  is a subspace.
- ♡ 2.2.24. Let  $W \subset V$  be a subspace. A subspace  $Z \subset V$  is called a *complementary subspace* to  $W$  if (i)  $W \cap Z = \{0\}$ , and (ii)  $W + Z = V$ , i.e., every  $\mathbf{v} \in V$  can be written as  $\mathbf{v} = \mathbf{w} + \mathbf{z}$  for  $\mathbf{w} \in W$  and  $\mathbf{z} \in Z$ . (a) Show that the  $x$ - and  $y$ -axes are complementary subspaces of  $\mathbb{R}^2$ . (b) Show that the lines  $x = y$  and  $x = 3y$  are complementary subspaces of  $\mathbb{R}^2$ . (c) Show that the line  $(a, 2a, 3a)^T$  and the plane  $x + 2y + 3z = 0$  are complementary subspaces of  $\mathbb{R}^3$ . (d) Prove that if  $\mathbf{v} = \mathbf{w} + \mathbf{z}$ , then  $\mathbf{w} \in W$  and  $\mathbf{z} \in Z$  are uniquely determined.
- 2.2.25. (a) Show that  $V_0 = \{(\mathbf{v}, \mathbf{0}) \mid \mathbf{v} \in V\}$  and  $W_0 = \{(\mathbf{0}, \mathbf{w}) \mid \mathbf{w} \in W\}$  are complementary subspaces, as in Exercise 2.2.24, of the Cartesian product space  $V \times W$ , as defined in Exercise 2.1.13. (b) Prove that the *diagonal*  $D = \{(\mathbf{v}, \mathbf{v})\}$  and the *anti-diagonal*  $A = \{(\mathbf{v}, -\mathbf{v})\}$  are complementary subspaces of  $V \times V$ .
- 2.2.26. Show that the set of skew-symmetric  $n \times n$  matrices forms a complementary subspace to the set of symmetric  $n \times n$  matrices. Explain why this implies that every square matrix can be uniquely written as the sum of a symmetric and a skew-symmetric matrix.

2.2.27. (a) Show that the set of even functions,  $f(-x) = f(x)$ , is a subspace of the vector space of all functions  $\mathcal{F}(\mathbb{R})$ . (b) Show that the set of odd functions,  $g(-x) = -g(x)$ , forms a complementary subspace, as defined in Exercise 2.2.24. (c) Explain why every function can be uniquely written as the sum of an even function and an odd function.

◇ 2.2.28. Let  $V$  be a vector space. A subset of the form  $A = \{ \mathbf{w} + \mathbf{a} \mid \mathbf{w} \in W \}$ , where  $W \subset V$  is a subspace and  $\mathbf{a} \in V$  is a fixed vector, is known as an *affine subspace* of  $V$ . (a) Show that an affine subspace  $A \subset V$  is a genuine subspace if and only if  $\mathbf{a} \in W$ . (b) Draw the affine subspaces  $A \subset \mathbb{R}^2$  when (i)  $W$  is the  $x$ -axis and  $\mathbf{a} = (2, 1)^T$ , (ii)  $W$  is the line  $y = \frac{3}{2}x$  and  $\mathbf{a} = (1, 1)^T$ , (iii)  $W$  is the line  $\{ (t, -t)^T \mid t \in \mathbb{R} \}$ , and  $\mathbf{a} = (2, -2)^T$ . (c) Prove that every affine subspace  $A \subset \mathbb{R}^2$  is either a point, a line, or all of  $\mathbb{R}^2$ . (d) Show that the plane  $x - 2y + 3z = 1$  is an affine subspace of  $\mathbb{R}^3$ . (e) Show that the set of all polynomials such that  $p(0) = 1$  is an affine subspace of  $\mathcal{P}^{(n)}$ .

◇ 2.2.29. *Quotient spaces*: Let  $V$  be a vector space and  $W \subset V$  a subspace. We say that two vectors  $\mathbf{u}, \mathbf{v} \in V$  are *equivalent modulo  $W$*  if  $\mathbf{u} - \mathbf{v} \in W$ . (a) Show that this defines an *equivalence relation*, written  $\mathbf{u} \sim_W \mathbf{v}$  on  $V$ , i.e., (i)  $\mathbf{v} \sim_W \mathbf{v}$  for every  $\mathbf{v}$ ; (ii) if  $\mathbf{u} \sim_W \mathbf{v}$ , then  $\mathbf{v} \sim_W \mathbf{u}$ ; and (iii) if, in addition,  $\mathbf{v} \sim_W \mathbf{z}$ , then  $\mathbf{u} \sim_W \mathbf{z}$ . (b) The *equivalence class* of a vector  $\mathbf{u} \in V$  is defined as the set of all equivalent vectors, written  $[\mathbf{u}]_W = \{ \mathbf{v} \in V \mid \mathbf{v} \sim_W \mathbf{u} \}$ . Show that  $[\mathbf{0}]_W = W$ . (c) Let  $V = \mathbb{R}^2$  and  $W = \{ (x, y)^T \mid x = 2y \}$ . Sketch a picture of several equivalence classes as subsets of  $\mathbb{R}^2$ . (d) Show that each equivalence class  $[\mathbf{u}]_W$  for  $\mathbf{u} \in V$  is an affine subspace of  $V$ , as in Exercise 2.2.28. (e) Prove that the set of equivalence classes, called the *quotient space* and denoted by  $V/W = \{ [\mathbf{u}] \mid \mathbf{u} \in V \}$ , forms a vector space under the operations of addition,  $[\mathbf{u}]_W + [\mathbf{v}]_W = [\mathbf{u} + \mathbf{v}]_W$ , and scalar multiplication,  $c[\mathbf{u}]_W = [c\mathbf{u}]_W$ . What is the zero element? Thus, you first need to prove that these operations are well defined, and then demonstrate the vector space axioms.

◇ 2.2.30. Define  $f(x) = \begin{cases} e^{-1/x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$

(a) Prove that all derivatives of  $f$  vanish at the origin:  $f^{(n)}(0) = 0$  for  $n = 0, 1, 2, \dots$ .

(b) Prove that  $f(x)$  is not analytic by showing that its Taylor series at  $a = 0$  does not converge to  $f(x)$  when  $x > 0$ .

2.2.31. Let  $f(x) = \frac{1}{1+x^2}$ . (a) Find the Taylor series of  $f$  at  $a = 0$ . (b) Prove that the Taylor series converges for  $|x| < 1$ , but diverges for  $|x| \geq 1$ . (c) Prove that  $f(x)$  is analytic at  $x = 0$ .

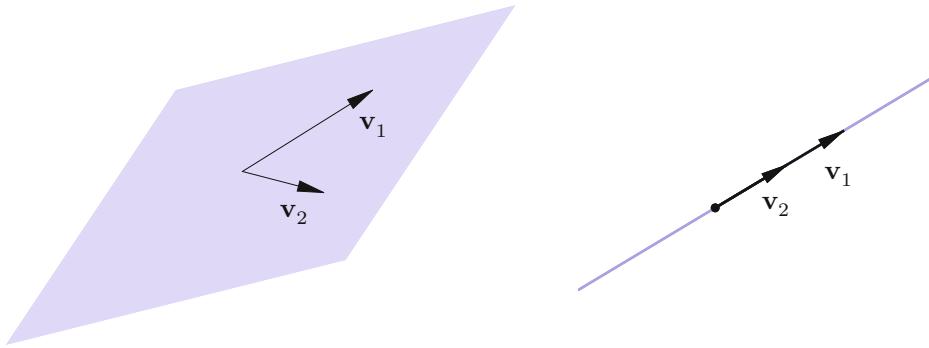
## 2.3 Span and Linear Independence

The definition of the span of a collection of elements of a vector space generalizes, in a natural fashion, the geometric notion of two vectors spanning a plane in  $\mathbb{R}^3$ . As such, it describes the first of two universal methods for constructing subspaces of vector spaces.

**Definition 2.13.** Let  $\mathbf{v}_1, \dots, \mathbf{v}_k$  be elements of a vector space  $V$ . A sum of the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_k\mathbf{v}_k = \sum_{i=1}^k c_i\mathbf{v}_i, \quad (2.5)$$

where the coefficients  $c_1, c_2, \dots, c_k$  are any scalars, is known as a *linear combination* of the elements  $\mathbf{v}_1, \dots, \mathbf{v}_k$ . Their *span* is the subset  $W = \text{span} \{ \mathbf{v}_1, \dots, \mathbf{v}_k \} \subset V$  consisting of all possible linear combinations with scalars  $c_1, \dots, c_k \in \mathbb{R}$ .



**Figure 2.4.** Plane and Line Spanned by Two Vectors.

For instance,  $3\mathbf{v}_1 + \mathbf{v}_2 - 2\mathbf{v}_3$ ,  $8\mathbf{v}_1 - \frac{1}{3}\mathbf{v}_3 = 8\mathbf{v}_1 + 0\mathbf{v}_2 - \frac{1}{3}\mathbf{v}_3$ ,  $\mathbf{v}_2 = 0\mathbf{v}_1 + 1\mathbf{v}_2 + 0\mathbf{v}_3$ , and  $\mathbf{0} = 0\mathbf{v}_1 + 0\mathbf{v}_2 + 0\mathbf{v}_3$  are four different linear combinations of the three vector space elements  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V$ .

The key observation is that the span always forms a subspace.

**Proposition 2.14.** The span  $W = \text{span} \{ \mathbf{v}_1, \dots, \mathbf{v}_k \}$  of any finite collection of vector space elements  $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$  is a subspace of the underlying vector space  $V$ .

*Proof:* We need to show that if

$$\mathbf{v} = c_1\mathbf{v}_1 + \dots + c_k\mathbf{v}_k \quad \text{and} \quad \widehat{\mathbf{v}} = \widehat{c}_1\mathbf{v}_1 + \dots + \widehat{c}_k\mathbf{v}_k$$

are any two linear combinations, then their sum is also a linear combination, since

$$\mathbf{v} + \widehat{\mathbf{v}} = (c_1 + \widehat{c}_1)\mathbf{v}_1 + \dots + (c_k + \widehat{c}_k)\mathbf{v}_k = \tilde{c}_1\mathbf{v}_1 + \dots + \tilde{c}_k\mathbf{v}_k,$$

where  $\tilde{c}_i = c_i + \widehat{c}_i$ . Similarly, for any scalar multiple,

$$a\mathbf{v} = (a c_1)\mathbf{v}_1 + \dots + (a c_k)\mathbf{v}_k = c_1^*\mathbf{v}_1 + \dots + c_k^*\mathbf{v}_k,$$

where  $c_i^* = a c_i$ , which completes the proof. *Q.E.D.*

**Example 2.15.** Examples of subspaces spanned by vectors in  $\mathbb{R}^3$ :

- (i) If  $\mathbf{v}_1 \neq \mathbf{0}$  is any non-zero vector in  $\mathbb{R}^3$ , then its span is the line  $\{ c\mathbf{v}_1 \mid c \in \mathbb{R} \}$  consisting of all vectors *parallel* to  $\mathbf{v}_1$ . If  $\mathbf{v}_1 = \mathbf{0}$ , then its span just contains the origin.
- (ii) If  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are any two vectors in  $\mathbb{R}^3$ , then their span is the set of all vectors of the form  $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$ . Typically, such a span prescribes a plane passing through the origin. However, if  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are parallel, then their span is just a line. The most degenerate case occurs when  $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{0}$ , where the span is just a point — the origin.
- (iii) If we are given three non-coplanar vectors  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ , then their span is all of  $\mathbb{R}^3$ , as we shall prove below. However, if they all lie in a plane, then their span is the plane — unless they are all parallel, in which case their span is a line — or, in the completely degenerate situation  $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{v}_3 = \mathbf{0}$ , a single point.

Thus, every subspace of  $\mathbb{R}^3$  can be realized as the span of some set of vectors. One can consider subspaces spanned by four or more vectors in  $\mathbb{R}^3$ , but these continue to be limited to being either a point (the origin), a line, a plane, or the entire three-dimensional space.

A crucial question is to determine when a given vector belongs to the span of a prescribed collection.

**Example 2.16.** Let  $W \subset \mathbb{R}^3$  be the plane spanned by the vectors  $\mathbf{v}_1 = (1, -2, 1)^T$  and  $\mathbf{v}_2 = (2, -3, 1)^T$ . Question: Is the vector  $\mathbf{v} = (0, 1, -1)^T$  an element of  $W$ ? To answer, we need to see whether we can find scalars  $c_1, c_2$  such that

$$\mathbf{v} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2; \quad \text{that is,} \quad \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 + 2c_2 \\ -2c_1 - 3c_2 \\ c_1 + c_2 \end{pmatrix}.$$

Thus,  $c_1, c_2$  must satisfy the linear algebraic system

$$c_1 + 2c_2 = 0, \quad -2c_1 - 3c_2 = 1, \quad c_1 + c_2 = -1.$$

Applying Gaussian Elimination, we find the solution  $c_1 = -2$ ,  $c_2 = 1$ , and so  $\mathbf{v} = -2\mathbf{v}_1 + \mathbf{v}_2$  does belong to the span. On the other hand,  $\tilde{\mathbf{v}} = (1, 0, 0)^T$  does not belong to  $W$ . Indeed, there are no scalars  $c_1, c_2$  such that  $\tilde{\mathbf{v}} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$ , because the corresponding linear system is incompatible.

**Warning.** It is entirely possible for different sets of vectors to span the *same* subspace. For instance,  $\mathbf{e}_1 = (1, 0, 0)^T$  and  $\mathbf{e}_2 = (0, 1, 0)^T$  span the  $xy$ -plane in  $\mathbb{R}^3$ , as do the three coplanar vectors  $\mathbf{v}_1 = (1, -1, 0)^T$ ,  $\mathbf{v}_2 = (-1, 2, 0)^T$ ,  $\mathbf{v}_3 = (2, 1, 0)^T$ .

**Example 2.17.** Let  $V = \mathcal{F}(\mathbb{R})$  denote the space of all scalar functions  $f(x)$ .

(a) The span of the three monomials  $f_1(x) = 1$ ,  $f_2(x) = x$ , and  $f_3(x) = x^2$  is the set of all functions of the form

$$f(x) = c_1 f_1(x) + c_2 f_2(x) + c_3 f_3(x) = c_1 + c_2 x + c_3 x^2,$$

where  $c_1, c_2, c_3$  are arbitrary scalars (constants). In other words,  $\text{span} \{1, x, x^2\} = \mathcal{P}^{(2)}$  is the subspace of all quadratic (degree  $\leq 2$ ) polynomials. In a similar fashion, the space  $\mathcal{P}^{(n)}$  of polynomials of degree  $\leq n$  is spanned by the monomials  $1, x, x^2, \dots, x^n$ .

(b) The next example plays a key role in many applications. Let  $0 \neq \omega \in \mathbb{R}$ . Consider the two basic trigonometric functions  $f_1(x) = \cos \omega x$ ,  $f_2(x) = \sin \omega x$  of frequency  $\omega$ , and hence period  $2\pi/\omega$ . Their span consists of all functions of the form

$$f(x) = c_1 f_1(x) + c_2 f_2(x) = c_1 \cos \omega x + c_2 \sin \omega x. \quad (2.6)$$

For example, the function  $\cos(\omega x + 2)$  lies in the span because, by the addition formula for the cosine,

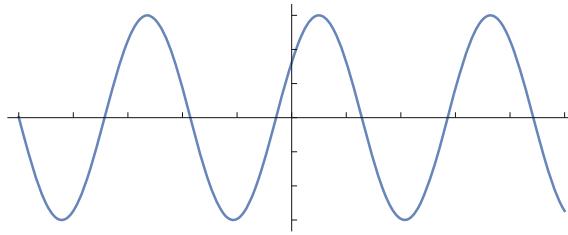
$$\cos(\omega x + 2) = (\cos 2) \cos \omega x - (\sin 2) \sin \omega x$$

is a linear combination of  $\cos \omega x$  and  $\sin \omega x$ , with respective coefficients  $\cos 2$ ,  $\sin 2$ . Indeed, we can express a general function in the span in the alternative *phase-amplitude form*

$$f(x) = c_1 \cos \omega x + c_2 \sin \omega x = r \cos(\omega x - \delta), \quad (2.7)$$

in which  $r \geq 0$  is known as the *amplitude* and  $0 \leq \delta < 2\pi$  the *phase shift*. Indeed, expanding the right-hand side, we obtain

$$r \cos(\omega x - \delta) = (r \cos \delta) \cos \omega x + (r \sin \delta) \sin \omega x, \quad \text{and hence } c_1 = r \cos \delta, \quad c_2 = r \sin \delta.$$



**Figure 2.5.** Graph of  $3 \cos(2x - 1)$ .

Thus,  $(r, \delta)$  are the polar coordinates of the point  $\mathbf{c} = (c_1, c_2) \in \mathbb{R}^2$  prescribed by the coefficients. We conclude that every linear combination of  $\sin \omega x$  and  $\cos \omega x$  can be rewritten as a single cosine containing an extra phase shift. [Figure 2.5](#) shows the particular function  $3 \cos(2x - 1)$ , which has amplitude  $r = 3$ , frequency  $\omega = 2$ , and phase shift  $\delta = 1$ . The first peak appears at  $x = \delta/\omega = \frac{1}{2}$ .

(c) The space  $\mathcal{T}^{(2)}$  of *quadratic trigonometric polynomials* is spanned by the functions

$$1, \quad \cos x, \quad \sin x, \quad \cos^2 x, \quad \cos x \sin x, \quad \sin^2 x.$$

Its general element is a linear combination

$$q(x) = c_0 + c_1 \cos x + c_2 \sin x + c_3 \cos^2 x + c_4 \cos x \sin x + c_5 \sin^2 x, \quad (2.8)$$

where  $c_0, \dots, c_5$  are arbitrary constants. A more useful spanning set for the same subspace consists of the trigonometric functions

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x. \quad (2.9)$$

Indeed, by the double-angle formulas, both

$$\cos 2x = \cos^2 x - \sin^2 x, \quad \sin 2x = 2 \sin x \cos x,$$

have the form of a quadratic trigonometric polynomial (2.8), and hence both belong to  $\mathcal{T}^{(2)}$ . On the other hand, we can write

$$\cos^2 x = \frac{1}{2} \cos 2x + \frac{1}{2}, \quad \cos x \sin x = \frac{1}{2} \sin 2x, \quad \sin^2 x = -\frac{1}{2} \cos 2x + \frac{1}{2},$$

in terms of the functions (2.9). Therefore, the original linear combination (2.8) can be written in the alternative form

$$\begin{aligned} q(x) &= (c_0 + \frac{1}{2}c_3 + \frac{1}{2}c_5) + c_1 \cos x + c_2 \sin x + (\frac{1}{2}c_3 - \frac{1}{2}c_5) \cos 2x + \frac{1}{2}c_4 \sin 2x \\ &= \hat{c}_0 + \hat{c}_1 \cos x + \hat{c}_2 \sin x + \hat{c}_3 \cos 2x + \hat{c}_4 \sin 2x, \end{aligned} \quad (2.10)$$

and so the functions (2.9) do indeed span  $\mathcal{T}^{(2)}$ . It is worth noting that we first characterized  $\mathcal{T}^{(2)}$  as the span of 6 functions, whereas the second characterization required only 5 functions. It turns out that 5 is the minimal number of functions needed to span  $\mathcal{T}^{(2)}$ , but the proof of this fact will be deferred until Chapter 4.

(d) The homogeneous linear ordinary differential equation

$$u'' + 2u' - 3u = 0 \quad (2.11)$$

considered in part (i) of Example 2.12 has two solutions:  $f_1(x) = e^x$  and  $f_2(x) = e^{-3x}$ . (Now may be a good time for you to review the basic techniques for solving linear, constant

coefficient ordinary differential equations, cf. [7, 22]; see also Chapter 7.) Its general solution is, in fact, a linear combination

$$u = c_1 f_1(x) + c_2 f_2(x) = c_1 e^x + c_2 e^{-3x},$$

where  $c_1, c_2$  are arbitrary scalars. Thus, the vector space of solutions to (2.11) is described as the span of these two basic solutions. The fact that there are no other solutions is not obvious, but relies on the basic uniqueness theorem for ordinary differential equations; further details can be found in Theorem 7.34.

**Remark.** One can also define the span of an infinite collection of elements of a vector space. To avoid convergence issues, one should consider only finite linear combinations (2.5). For example, the span of the monomials  $1, x, x^2, x^3, \dots$  is the subspace  $\mathcal{P}^{(\infty)}$  of all polynomials — *not* the space of analytic functions or convergent Taylor series. Similarly, the span of the functions  $1, \cos x, \sin x, \cos 2x, \sin 2x, \cos 3x, \sin 3x, \dots$  is the space  $\mathcal{T}^{(\infty)}$  containing all *trigonometric polynomials*, of fundamental importance in the theory of Fourier series, [61].

---

## Exercises

2.3.1. Show that  $\begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}$  belongs to the subspace of  $\mathbb{R}^3$  spanned by  $\begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ -4 \\ 1 \end{pmatrix}$  by writing

it as a linear combination of the spanning vectors.

2.3.2. Show that  $\begin{pmatrix} -3 \\ 7 \\ 6 \\ 1 \end{pmatrix}$  is in the subspace of  $\mathbb{R}^4$  spanned by  $\begin{pmatrix} 1 \\ -3 \\ -2 \\ 0 \end{pmatrix}, \begin{pmatrix} -2 \\ 6 \\ 3 \\ 4 \end{pmatrix}$  and  $\begin{pmatrix} -2 \\ 4 \\ 6 \\ -7 \end{pmatrix}$ .

2.3.3. (a) Determine whether  $\begin{pmatrix} 1 \\ -2 \\ -3 \end{pmatrix}$  is in the span of  $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$  and  $\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ . (b) Is  $\begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}$  in the span of  $\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} -1 \\ -2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}$ ? (c) Is  $\begin{pmatrix} 3 \\ 0 \\ -1 \\ -2 \end{pmatrix}$  in the span of  $\begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \\ -1 \end{pmatrix}$ ?

2.3.4. Which of the following sets of vectors span all of  $\mathbb{R}^2$ ? (a)  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ; (b)  $\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ ; (c)  $\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ ; (d)  $\begin{pmatrix} 6 \\ -9 \end{pmatrix}, \begin{pmatrix} -4 \\ 6 \end{pmatrix}$ ; (e)  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ ; (f)  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}$ .

2.3.5. (a) Graph the subspace of  $\mathbb{R}^3$  spanned by the vector  $\mathbf{v}_1 = (3, 0, 1)^T$ .

(b) Graph the subspace spanned by the vectors  $\mathbf{v}_1 = (3, -2, -1)^T, \mathbf{v}_2 = (-2, 0, -1)^T$ .

(c) Graph the span of  $\mathbf{v}_1 = (1, 0, -1)^T, \mathbf{v}_2 = (0, -1, 1)^T, \mathbf{v}_3 = (1, -1, 0)^T$ .

2.3.6. Let  $U$  be the subspace of  $\mathbb{R}^3$  spanned by  $\mathbf{u}_1 = (1, 2, 3)^T, \mathbf{u}_2 = (2, -1, 0)^T$ . Let  $V$  be the subspace spanned by  $\mathbf{v}_1 = (5, 0, 3)^T, \mathbf{v}_2 = (3, 1, 3)^T$ . Is  $V$  a subspace of  $U$ ? Are  $U$  and  $V$  the same?

2.3.7. (a) Let  $S$  be the subspace of  $\mathcal{M}_{2 \times 2}$  consisting of all symmetric  $2 \times 2$  matrices. Show that  $S$  is spanned by the matrices  $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ , and  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . (b) Find a spanning set of the space of symmetric  $3 \times 3$  matrices.

2.3.8. (a) Determine whether the polynomials  $x^2 + 1, x^2 - 1, x^2 + x + 1$  span  $\mathcal{P}^{(2)}$ .

(b) Do  $x^3 - 1, x^2 + 1, x - 1, 1$  span  $\mathcal{P}^{(3)}$ ? (c) What about  $x^3, x^2 + 1, x^2 - x, x + 1$ ?

2.3.9. Determine whether any of the following functions lies in the subspace spanned by  $1, x, \sin x, \sin^2 x$ : (a)  $3 - 5x$ , (b)  $x^2 + \sin^2 x$ , (c)  $\sin x - 2 \cos x$ , (d)  $\cos^2 x$ , (e)  $x \sin x$ , (f)  $e^x$ .

2.3.10. Write the following trigonometric functions in phase-amplitude form:

(a)  $\sin 3x$ , (b)  $\cos x - \sin x$ , (c)  $3 \cos 2x + 4 \sin 2x$ , (d)  $\cos x \sin x$ .

2.3.11. (a) Prove that the set of solutions to the homogeneous ordinary differential equation  $u'' - 4u' + 3u = 0$  is a vector space. (b) Write the solution space as the span of a finite number of functions. (c) What is the minimal number of functions needed to span the solution space?

2.3.12. Explain why the functions  $1, \cos x, \sin x$  span the solution space to the third order ordinary differential equation  $u''' + u' = 0$ .

2.3.13. Find a finite set of real functions that spans the solution space to the following homogeneous ordinary differential equations: (a)  $u' - 2u = 0$ , (b)  $u'' + 4u = 0$ , (c)  $u'' - 3u' = 0$ , (d)  $u'' + u' + u = 0$ , (e)  $u''' - 5u'' = 0$ , (f)  $u^{(4)} + u = 0$ .

2.3.14. Consider the boundary value problem  $u'' + 4u = 0$ ,  $0 \leq x \leq \pi$ ,  $u(0) = 0$ ,  $u(\pi) = 0$ .

(a) Prove, without solving, that the set of solutions forms a vector space.

(b) Write this space as the span of one or more functions. *Hint:* First solve the differential equation; then find out which solutions satisfy the boundary conditions.

2.3.15. Which of the following functions lie in the span of the vector-valued functions

$$\mathbf{f}_1(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}, \quad \mathbf{f}_2(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}, \quad \mathbf{f}_3(x) = \begin{pmatrix} x \\ 2x \end{pmatrix};$$

(a)  $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1-2x \\ 1-x \end{pmatrix}$ , (c)  $\begin{pmatrix} 1-2x \\ -1-x \end{pmatrix}$ , (d)  $\begin{pmatrix} 1+x^2 \\ 1-x^2 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2-x \\ 0 \end{pmatrix}$ .

2.3.16. *True or false:* The zero vector belongs to the span of any collection of vectors.

2.3.17. Prove or give a counter-example: if  $\mathbf{z}$  is a linear combination of  $\mathbf{u}, \mathbf{v}, \mathbf{w}$ , then  $\mathbf{w}$  is a linear combination of  $\mathbf{u}, \mathbf{v}, \mathbf{z}$ .

◇ 2.3.18. Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_m$  span  $V$ . Let  $\mathbf{v}_{m+1}, \dots, \mathbf{v}_n \in V$  be any other elements. Prove that the combined collection  $\mathbf{v}_1, \dots, \mathbf{v}_n$  also spans  $V$ .

◇ 2.3.19. (a) Show that if  $\mathbf{v}$  is a linear combination of  $\mathbf{v}_1, \dots, \mathbf{v}_m$ , and each  $\mathbf{v}_j$  is a linear combination of  $\mathbf{w}_1, \dots, \mathbf{w}_n$ , then  $\mathbf{v}$  is a linear combination of  $\mathbf{w}_1, \dots, \mathbf{w}_n$ .

(b) Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_m$  span  $V$ . Let  $\mathbf{w}_1, \dots, \mathbf{w}_m \in V$  be any other elements. Suppose that each  $\mathbf{v}_i$  can be written as a linear combination of  $\mathbf{w}_1, \dots, \mathbf{w}_m$ . Prove that  $\mathbf{w}_1, \dots, \mathbf{w}_m$  also span  $V$ .

◇ 2.3.20. The span of an infinite collection  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots \in V$  of vector space elements is defined as the set of all *finite* linear combinations  $\sum_{i=1}^n c_i \mathbf{v}_i$ , where  $n < \infty$  is finite but arbitrary.

(a) Prove that the span defines a subspace of the vector space  $V$ .

(b) What is the span of the monomials  $1, x, x^2, x^3, \dots$ ?

## Linear Independence and Dependence

Most of the time, all of the vectors used to form a span are essential. For example, we cannot use fewer than two vectors to span a plane in  $\mathbb{R}^3$ , since the span of a single vector is at most a line. However, in degenerate situations, some of the spanning elements may be

redundant. For instance, if the two vectors are parallel, then their span is a line, but only one of the vectors is really needed to prescribe the line. Similarly, the subspace spanned by the polynomials  $p_1(x) = x - 2$ ,  $p_2(x) = 3x + 4$ ,  $p_3(x) = -x + 1$ , is the vector space  $\mathcal{P}^{(1)}$  consisting of all linear polynomials. But only two of the polynomials are really required to span  $\mathcal{P}^{(1)}$ . (The reason will become clear soon, but you may wish to see whether you can demonstrate this on your own.) The elimination of such superfluous spanning elements is encapsulated in the following important definition.

**Definition 2.18.** The vector space elements  $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$  are called *linearly dependent* if there exist scalars  $c_1, \dots, c_k$ , *not all zero*, such that

$$c_1\mathbf{v}_1 + \dots + c_k\mathbf{v}_k = \mathbf{0}. \quad (2.12)$$

Elements that are not linearly dependent are called *linearly independent*.

The restriction that not all the  $c_i$ 's are zero is essential: if  $c_1 = \dots = c_k = 0$ , then the linear combination (2.12) is automatically zero. Thus, to check linear independence, one needs to show that the *only* linear combination that produces the zero vector (2.12) is this trivial one. In other words,  $c_1 = \dots = c_k = 0$  is the *one and only* solution to the vector equation (2.12).

**Example 2.19.** Some examples of linear independence and dependence:

(a) The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 4 \\ 3 \end{pmatrix},$$

are linearly dependent, because

$$\mathbf{v}_1 - 2\mathbf{v}_2 + \mathbf{v}_3 = \mathbf{0}.$$

On the other hand, the first two vectors  $\mathbf{v}_1, \mathbf{v}_2$  are linearly independent. To see this, suppose that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \begin{pmatrix} c_1 \\ 2c_1 + 3c_2 \\ -c_1 + c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

For this to happen,  $c_1, c_2$  must satisfy the homogeneous linear system

$$c_1 = 0, \quad 2c_1 + 3c_2 = 0, \quad -c_1 + c_2 = 0,$$

which, as you can check, has only the trivial solution  $c_1 = c_2 = 0$ .

(b) In general, any collection  $\mathbf{v}_1, \dots, \mathbf{v}_k$  that includes the zero vector, say  $\mathbf{v}_1 = \mathbf{0}$ , is automatically linearly dependent, since  $1\mathbf{0} + 0\mathbf{v}_2 + \dots + 0\mathbf{v}_k = \mathbf{0}$  is a nontrivial linear combination that adds up to  $\mathbf{0}$ .

(c) Two vectors  $\mathbf{v}, \mathbf{w} \in V$  are linearly dependent if and only if they are *parallel*, meaning that one is a scalar multiple of the other. Indeed, if  $\mathbf{v} = a\mathbf{w}$ , then  $\mathbf{v} - a\mathbf{w} = \mathbf{0}$  is a nontrivial linear combination summing to zero. Conversely, if  $c\mathbf{v} + d\mathbf{w} = \mathbf{0}$  and  $c \neq 0$ , then  $\mathbf{v} = -(d/c)\mathbf{w}$ , while if  $c = 0$  but  $d \neq 0$ , then  $\mathbf{w} = \mathbf{0}$ .

(d) The polynomials

$$p_1(x) = x - 2, \quad p_2(x) = x^2 - 5x + 4, \quad p_3(x) = 3x^2 - 4x, \quad p_4(x) = x^2 - 1,$$

are linearly dependent, since

$$p_1(x) + p_2(x) - p_3(x) + 2p_4(x) \equiv 0$$

is a nontrivial linear combination that vanishes identically. On the other hand, the first three polynomials,

$$p_1(x) = x - 2, \quad p_2(x) = x^2 - 5x + 4, \quad p_3(x) = 3x^2 - 4x,$$

are linearly independent. Indeed, if the linear combination

$$c_1 p_1(x) + c_2 p_2(x) + c_3 p_3(x) = (c_2 + 3c_3)x^2 + (c_1 - 5c_2 - 4c_3)x - 2c_1 + 4c_2 \equiv 0$$

is the zero polynomial, then its coefficients must vanish, and hence  $c_1, c_2, c_3$  are required to solve the homogeneous linear system

$$c_2 + 3c_3 = 0, \quad c_1 - 5c_2 - 4c_3 = 0, \quad -2c_1 + 4c_2 = 0.$$

But this has only the trivial solution  $c_1 = c_2 = c_3 = 0$ , and so linear independence follows.

**Remark.** In the last example, we are using the basic fact that a polynomial is identically zero,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \equiv 0 \quad \text{for all } x,$$

if and only if its coefficients all vanish:  $a_0 = a_1 = \cdots = a_n = 0$ . This is equivalent to the “obvious” fact that the basic monomial functions  $1, x, x^2, \dots, x^n$  are linearly independent. Exercise 2.3.36 asks for a bona fide proof.

**Example 2.20.** The trigonometric functions

$$1, \quad \cos x, \quad \sin x, \quad \cos^2 x, \quad \cos x \sin x, \quad \sin^2 x,$$

which were used to define the vector space  $\mathcal{T}^{(2)}$  of quadratic trigonometric polynomials, are, in fact, linearly dependent. This is a consequence of the basic trigonometric identity

$$\cos^2 x + \sin^2 x \equiv 1,$$

which can be rewritten as a nontrivial linear combination

$$1 + 0 \cos x + 0 \sin x + (-1) \cos^2 x + 0 \cos x \sin x + (-1) \sin^2 x \equiv 0$$

that equals the zero function. On the other hand, the alternative spanning set

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x$$

is linearly independent, since the only identically zero linear combination,

$$c_0 + c_1 \cos x + c_2 \sin x + c_3 \cos 2x + c_4 \sin 2x \equiv 0,$$

turns out to be the trivial one  $c_0 = \cdots = c_4 = 0$ . However, the latter fact is not as obvious, and requires a bit of work to prove directly; see Exercise 2.3.37. An easier proof, based on orthogonality, will appear in Chapter 4.

Let us now focus our attention on the linear independence or dependence of a set of vectors  $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$  in Euclidean space. We begin by forming the  $n \times k$  matrix  $A = (\mathbf{v}_1 \dots \mathbf{v}_k)$  whose *columns* are the given vectors. (The fact that we use column vectors is essential here.) Our analysis is based on the very useful formula

$$A\mathbf{c} = c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k, \quad \text{where} \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix}, \quad (2.13)$$

that expresses any linear combination in terms of matrix multiplication. For example,

$$\begin{pmatrix} 1 & 3 & 0 \\ -1 & 2 & 1 \\ 4 & -1 & -2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} c_1 + 3c_2 \\ -c_1 + 2c_2 + c_3 \\ 4c_1 - c_2 - 2c_3 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ -1 \\ 4 \end{pmatrix} + c_2 \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} + c_3 \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}.$$

Formula (2.13) follows directly from the rules of matrix multiplication; see also Exercise 1.2.34(c). It enables us to reformulate the notions of linear independence and span of vectors in  $\mathbb{R}^n$  in terms of linear algebraic systems. The key result is the following:

**Theorem 2.21.** Let  $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$  and let  $A = (\mathbf{v}_1 \dots \mathbf{v}_k)$  be the corresponding  $n \times k$  matrix whose columns are the given vectors.

- (a) The vectors  $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$  are linearly dependent if and only if there is a non-zero solution  $\mathbf{c} \neq \mathbf{0}$  to the homogeneous linear system  $A\mathbf{c} = \mathbf{0}$ .
- (b) The vectors are linearly independent if and only if the only solution to the homogeneous system  $A\mathbf{c} = \mathbf{0}$  is the trivial one,  $\mathbf{c} = \mathbf{0}$ .
- (c) A vector  $\mathbf{b}$  lies in the span of  $\mathbf{v}_1, \dots, \mathbf{v}_k$  if and only if the linear system  $A\mathbf{c} = \mathbf{b}$  is compatible, i.e., has at least one solution.

*Proof:* We prove the first statement, leaving the other two as exercises for the reader. The condition that  $\mathbf{v}_1, \dots, \mathbf{v}_k$  be linearly dependent is that there exists a nonzero vector

$$\mathbf{c} = (c_1, c_2, \dots, c_k)^T \neq \mathbf{0} \quad \text{such that} \quad A\mathbf{c} = c_1\mathbf{v}_1 + \dots + c_k\mathbf{v}_k = \mathbf{0}.$$

Therefore, linear dependence requires the existence of a nontrivial solution to the homogeneous linear system  $A\mathbf{c} = \mathbf{0}$ . *Q.E.D.*

**Example 2.22.** Let us determine whether the vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -4 \\ 6 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 4 \\ 2 \\ 3 \end{pmatrix} \quad (2.14)$$

are linearly independent or linearly dependent. We combine them as column vectors into a single matrix

$$A = \begin{pmatrix} 1 & 3 & 1 & 4 \\ 2 & 0 & -4 & 2 \\ -1 & 4 & 6 & 3 \end{pmatrix}.$$

According to Theorem 2.21, we need to figure out whether there are any nontrivial solutions to the homogeneous equation  $A\mathbf{c} = \mathbf{0}$ ; this can be done by reducing  $A$  to row echelon form

$$U = \begin{pmatrix} 1 & 3 & 1 & 4 \\ 0 & -6 & -6 & -6 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (2.15)$$

The general solution to the homogeneous system  $A\mathbf{c} = \mathbf{0}$  is  $\mathbf{c} = (2c_3 - c_4, -c_3 - c_4, c_3, c_4)^T$ , where  $c_3, c_4$  — the free variables — are arbitrary. Any nonzero choice of  $c_3, c_4$  will produce a nontrivial linear combination

$$(2c_3 - c_4)\mathbf{v}_1 + (-c_3 - c_4)\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 = \mathbf{0}$$

that adds up to the zero vector. We conclude that the vectors (2.14) are linearly dependent.

In fact, in this particular case, we didn't even need to complete the row reduction if we only need to check linear (in)dependence. According to Theorem 1.47, any coefficient matrix with more columns than rows automatically has a nontrivial solution to the associated homogeneous system. This implies the following result:

**Lemma 2.23.** Any collection of  $k > n$  vectors in  $\mathbb{R}^n$  is linearly dependent.

**Warning.** The converse to this lemma is *not* true. For example,  $\mathbf{v}_1 = (1, 2, 3)^T$  and  $\mathbf{v}_2 = (-2, -4, -6)^T$  are two linearly dependent vectors in  $\mathbb{R}^3$ , since  $2\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{0}$ . For a collection of  $n$  or fewer vectors in  $\mathbb{R}^n$ , one needs to analyze the homogeneous linear system.

Lemma 2.23 is a particular case of the following general characterization of linearly independent vectors.

**Proposition 2.24.** A set of  $k$  vectors in  $\mathbb{R}^n$  is linearly independent if and only if the corresponding  $n \times k$  matrix  $A$  has rank  $k$ . In particular, this requires  $k \leq n$ .

Or, to state the result another way, the vectors are linearly independent if and only if the homogeneous linear system  $A\mathbf{c} = \mathbf{0}$  has no free variables. Proposition 2.24 is an immediate corollary of Theorems 2.21 and 1.47.

**Example 2.22 (continued).** Let us now see which vectors  $\mathbf{b} \in \mathbb{R}^3$  lie in the span of the vectors (2.14). According to Theorem 2.21, this will be the case if and only if the linear system  $A\mathbf{c} = \mathbf{b}$  has a solution. Since the resulting row echelon form (2.15) has a row of all zeros, there will be a compatibility condition on the entries of  $\mathbf{b}$ , and hence not every vector lies in the span. To find the precise condition, we augment the coefficient matrix, and apply the same row operations, leading to the reduced augmented matrix

$$\left( \begin{array}{cccc|c} 1 & 3 & 1 & 4 & b_1 \\ 0 & -6 & -6 & -6 & b_2 - 2b_1 \\ 0 & 0 & 0 & 0 & b_3 + \frac{7}{6}b_2 - \frac{4}{3}b_1 \end{array} \right).$$

Therefore,  $\mathbf{b} = (b_1, b_2, b_3)^T$  lies in the span if and only if  $-\frac{4}{3}b_1 + \frac{7}{6}b_2 + b_3 = 0$ . Thus, these four vectors span only a plane in  $\mathbb{R}^3$ .

The same method demonstrates that a collection of vectors will span all of  $\mathbb{R}^n$  if and only if the row echelon form of the associated matrix contains no all-zero rows, or, equivalently, the rank is equal to  $n$ , the number of rows in the matrix.

**Proposition 2.25.** A collection of  $k$  vectors spans  $\mathbb{R}^n$  if and only if their  $n \times k$  matrix has rank  $n$ . In particular, this requires  $k \geq n$ .

**Warning.** Not every collection of  $n$  or more vectors in  $\mathbb{R}^n$  will span all of  $\mathbb{R}^n$ . A counterexample was already provided by the vectors (2.14).

## Exercises

2.3.21. Determine whether the given vectors are linearly independent or linearly dependent:

(a)  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} -2 \\ -6 \end{pmatrix}$ , (c)  $\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 2 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \end{pmatrix}$ ,

(e)  $\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ , (f)  $\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ ,

(g)  $\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} -6 \\ -3 \end{pmatrix}$ , (h)  $\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ , (i)  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}$ ,

2.3.22. (a) Show that the vectors  $\begin{pmatrix} 1 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \\ 1 \\ -1 \end{pmatrix}$  are linearly independent. (b) Which

of the following vectors are in their span? (i)  $\begin{pmatrix} 1 \\ 1 \\ 2 \\ 1 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ , (iv)  $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ .

(c) Suppose  $\mathbf{b} = (a, b, c, d)^T$  lies in their span. What conditions must  $a, b, c, d$  satisfy?

2.3.23. (a) Show that the vectors  $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \\ -1 \end{pmatrix}$  are linearly independent.

(b) Show that they also span  $\mathbb{R}^4$ . (c) Write  $(1, 0, 0, 1)^T$  as a linear combination of them.

2.3.24. Determine whether the given row vectors are linearly independent or linearly dependent:

(a)  $(2, 1), (-1, 3), (5, 2)$ , (b)  $(1, 2, -1), (2, 4, -2)$ , (c)  $(1, 2, 3), (1, 4, 8), (1, 5, 7)$ ,  
 (d)  $(1, 1, 0), (1, 0, 3), (2, 2, 1)$ , (e)  $(1, 3, 4), (1, 2, 0, 3), (-3, -1, 2, -2), (3, -4, -4, 5)$ ,  
 (f)  $(2, 1, -1, 3), (-1, 3, 1, 0), (5, 1, 2, -3)$ .

2.3.25. *True or false:* The six  $3 \times 3$  permutation matrices (1.30) are linearly independent.

2.3.26. *True or false:* A set of vectors is linearly dependent if the zero vector belongs to their span.

2.3.27. Does a single vector ever define a linearly dependent set?

2.3.28. Let  $\mathbf{x}$  and  $\mathbf{y}$  be linearly independent elements of a vector space  $V$ . Show that  
 $\mathbf{u} = a\mathbf{x} + b\mathbf{y}$ , and  $\mathbf{v} = c\mathbf{x} + d\mathbf{y}$  are linearly independent if and only if  $ad - bc \neq 0$ . Is the entire collection  $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}$  linearly independent?

2.3.29. Prove or give a counterexample to the following statement: If  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are elements of a vector space  $V$  that do not span  $V$ , then  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are linearly independent.

◇ 2.3.30. Prove parts (b) and (c) of Theorem 2.21.

◇ 2.3.31. (a) Prove that if  $\mathbf{v}_1, \dots, \mathbf{v}_m$  are linearly independent, then every subset, e.g.,  $\mathbf{v}_1, \dots, \mathbf{v}_k$  with  $k < m$ , is also linearly independent. (b) Does the same hold true for linearly dependent vectors? Prove or give a counterexample.

2.3.32. (a) Determine whether the polynomials  $f_1(x) = x^2 - 3$ ,  $f_2(x) = 2 - x$ ,  $f_3(x) = (x - 1)^2$ , are linearly independent or linearly dependent.

(b) Do they span the vector space of all quadratic polynomials?

2.3.33. Determine whether the given functions are linearly independent or linearly dependent:

- (a)  $2 - x^2, 3x, x^2 + x - 2$ , (b)  $3x - 1, x(2x + 1), x(x - 1)$ ; (c)  $e^x, e^{x+1}$ ; (d)  $\sin x, \sin(x + 1)$ ; (e)  $e^x, e^{x+1}, e^{x+2}$ ; (f)  $\sin x, \sin(x + 1), \sin(x + 2)$ ; (g)  $e^x, xe^x, x^2e^x$ ; (h)  $e^x, e^{2x}, e^{3x}$ ; (i)  $x + y, x - y + 1, x + 3y + 2$  — these are functions of two variables.

2.3.34. Show that the functions  $f(x) = x$  and  $g(x) = |x|$  are linearly independent when considered as functions on all of  $\mathbb{R}$ , but are linearly dependent when considered as functions defined only on  $\mathbb{R}^+ = \{x > 0\}$ .

◇ 2.3.35. (a) Prove that the polynomials  $p_i(x) = \sum_{j=0}^n a_{ij} x^j$  for  $i = 1, \dots, k$  are linearly

independent if and only if the  $k \times (n + 1)$  matrix  $A$  whose entries are their coefficients  $a_{ij}$ ,  $1 \leq i \leq k$ ,  $0 \leq j \leq n$ , has rank  $k$ . (b) Formulate a similar matrix condition for testing whether another polynomial  $q(x)$  lies in their span. (c) Use (a) to determine whether  $p_1(x) = x^3 - 1$ ,  $p_2(x) = x^3 - 2x + 4$ ,  $p_3(x) = x^4 - 4x$ ,  $p_4(x) = x^2 + 1$ ,  $p_5(x) = -x^4 + 4x^3 + 2x + 1$  are linearly independent or linearly dependent. (d) Does the polynomial  $q(x) = x^3$  lie in their span? If so find a linear combination that adds up to  $q(x)$ .

◇ 2.3.36. The Fundamental Theorem of Algebra, [26], states that a non-zero polynomial of degree  $n$  has at most  $n$  distinct real roots, that is, real numbers  $x$  such that  $p(x) = 0$ . Use this fact to prove linear independence of the monomial functions  $1, x, x^2, \dots, x^n$ .

**Remark.** An elementary proof of the latter fact can be found in Exercise 5.5.38.

◇ 2.3.37. (a) Let  $x_1, x_2, \dots, x_n$  be a set of distinct sample points. Prove that the functions  $f_1(x), \dots, f_k(x)$  are linearly independent if their sample vectors  $\mathbf{f}_1, \dots, \mathbf{f}_k$  are linearly independent vectors in  $\mathbb{R}^n$ . (b) Give an example of linearly independent functions that have linearly dependent sample vectors. (c) Use this method to prove that the functions  $1, \cos x, \sin x, \cos 2x, \sin 2x$ , are linearly independent. *Hint:* You need at least 5 sample points.

2.3.38. Suppose  $\mathbf{f}_1(t), \dots, \mathbf{f}_k(t)$  are vector-valued functions from  $\mathbb{R}$  to  $\mathbb{R}^n$ . (a) Prove that if  $\mathbf{f}_1(t_0), \dots, \mathbf{f}_k(t_0)$  are linearly independent vectors in  $\mathbb{R}^n$  at one point  $t_0$ , then  $\mathbf{f}_1(t), \dots, \mathbf{f}_k(t)$  are linearly independent functions. (b) Show that  $\mathbf{f}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$  and  $\mathbf{f}_2(t) = \begin{pmatrix} 2t - 1 \\ 2t^2 - t \end{pmatrix}$  are linearly independent functions, even though at each  $t_0$ , the vectors  $\mathbf{f}_1(t_0), \mathbf{f}_2(t_0)$  are linearly dependent. Therefore, the converse to the result in part (a) is not valid.

◇ 2.3.39. The *Wronskian* of a pair of differentiable functions  $f(x), g(x)$  is the scalar function

$$W[f(x), g(x)] = \det \begin{pmatrix} f(x) & g(x) \\ f'(x) & g'(x) \end{pmatrix} = f(x)g'(x) - f'(x)g(x). \quad (2.16)$$

(a) Prove that if  $f, g$  are linearly dependent, then  $W[f(x), g(x)] \equiv 0$ . Hence, if  $W[f(x), g(x)] \not\equiv 0$ , then  $f, g$  are linearly independent. (b) Let  $f(x) = x^3$ ,  $g(x) = |x|^3$ . Prove that  $f, g \in C^2$  are twice continuously differentiable and linearly independent, but  $W[f(x), g(x)] \equiv 0$ . Thus, the Wronskian is *not* a fool-proof test for linear independence.

**Remark.** It can be proved, [7], that if  $f, g$  both satisfy a second order linear ordinary differential equation, then  $f, g$  are linearly dependent if and only if  $W[f(x), g(x)] \equiv 0$ .

## 2.4 Basis and Dimension

In order to span a vector space or subspace, we must employ a sufficient number of distinct elements. On the other hand, including too many elements in the spanning set will violate linear independence, and cause redundancies. The optimal spanning sets are those that are

also linearly independent. By combining the properties of span and linear independence, we arrive at the all-important concept of a “basis”.

**Definition 2.26.** A *basis* of a vector space  $V$  is a finite collection of elements  $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$  that (a) spans  $V$ , and (b) is linearly independent.

Bases are absolutely fundamental in all areas of linear algebra and linear analysis, including matrix algebra, Euclidean geometry, statistical analysis, solutions to linear differential equations — both ordinary and partial — linear boundary value problems, Fourier analysis, signal and image processing, data compression, control systems, and many others.

**Example 2.27.** The *standard basis* of  $\mathbb{R}^n$  consists of the  $n$  vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \dots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad (2.17)$$

so that  $\mathbf{e}_i$  is the vector with 1 in the  $i^{\text{th}}$  slot and 0's elsewhere. We already encountered these vectors — they are the columns of the  $n \times n$  identity matrix. They clearly span  $\mathbb{R}^n$ , since we can write any vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + \dots + x_n \mathbf{e}_n \quad (2.18)$$

as a linear combination, whose coefficients are its entries. Moreover, the only linear combination that yields the zero vector  $\mathbf{x} = \mathbf{0}$  is the trivial one  $x_1 = \dots = x_n = 0$ , which shows that  $\mathbf{e}_1, \dots, \mathbf{e}_n$  are linearly independent.

In the three-dimensional case  $\mathbb{R}^3$ , a common physical notation for the standard basis is

$$\mathbf{i} = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{j} = \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{k} = \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \quad (2.19)$$

This is but one of many possible bases for  $\mathbb{R}^3$ . Indeed, any three non-coplanar vectors can be used to form a basis. This is a consequence of the following general characterization of bases in Euclidean space as the columns of a nonsingular matrix.

**Theorem 2.28.** Every basis of  $\mathbb{R}^n$  consists of exactly  $n$  vectors. Furthermore, a set of  $n$  vectors  $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$  is a basis if and only if the  $n \times n$  matrix  $A = (\mathbf{v}_1 \dots \mathbf{v}_n)$  is nonsingular:  $\text{rank } A = n$ .

*Proof:* This is a direct consequence of Theorem 2.21. Linear independence requires that the only solution to the homogeneous system  $A\mathbf{c} = \mathbf{0}$  be the trivial one  $\mathbf{c} = \mathbf{0}$ . On the other hand, a vector  $\mathbf{b} \in \mathbb{R}^n$  will lie in the span of  $\mathbf{v}_1, \dots, \mathbf{v}_n$  if and only if the linear system  $A\mathbf{c} = \mathbf{b}$  has a solution. For  $\mathbf{v}_1, \dots, \mathbf{v}_n$  to span all of  $\mathbb{R}^n$ , this must hold for all possible right-hand sides  $\mathbf{b}$ . Theorem 1.7 tells us that both results require that  $A$  be nonsingular, i.e., have maximal rank  $n$ . *Q.E.D.*

Thus, every basis of  $n$ -dimensional Euclidean space  $\mathbb{R}^n$  contains the same number of vectors, namely  $n$ . This is a general fact, that motivates a linear algebraic characterization of dimension.

**Theorem 2.29.** Suppose the vector space  $V$  has a basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  for some  $n \in \mathbb{N}$ . Then every other basis of  $V$  has the same number,  $n$ , of elements in it. This number is called the *dimension* of  $V$ , and written  $\dim V = n$ .

The proof of Theorem 2.29 rests on the following lemma.

**Lemma 2.30.** Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n$  span a vector space  $V$ . Then every set of  $k > n$  elements  $\mathbf{w}_1, \dots, \mathbf{w}_k \in V$  is linearly dependent.

*Proof:* Let us write each element

$$\mathbf{w}_j = \sum_{i=1}^n a_{ij} \mathbf{v}_i, \quad j = 1, \dots, k,$$

as a linear combination of the spanning set. Then

$$c_1 \mathbf{w}_1 + \dots + c_k \mathbf{w}_k = \sum_{i=1}^n \sum_{j=1}^k a_{ij} c_j \mathbf{v}_i.$$

This linear combination will be zero whenever  $\mathbf{c} = (c_1, c_2, \dots, c_k)^T$  solves the homogeneous linear system

$$\sum_{j=1}^k a_{ij} c_j = 0, \quad i = 1, \dots, n,$$

consisting of  $n$  equations in  $k > n$  unknowns. Theorem 1.47 guarantees that every homogeneous system with more unknowns than equations always has a non-trivial solution  $\mathbf{c} \neq \mathbf{0}$ , and this immediately implies that  $\mathbf{w}_1, \dots, \mathbf{w}_k$  are linearly dependent. *Q.E.D.*

*Proof of Theorem 2.29:* Suppose we have two bases containing a different number of elements. By definition, the smaller basis spans the vector space. But then Lemma 2.30 tell us that the elements in the larger purported basis must be linearly dependent, which contradicts our initial assumption that the latter is a basis. *Q.E.D.*

As a direct consequence, we can now give a precise meaning to the optimality of bases.

**Theorem 2.31.** Suppose  $V$  is an  $n$ -dimensional vector space. Then

- (a) Every set of more than  $n$  elements of  $V$  is linearly dependent.
- (b) No set of fewer than  $n$  elements spans  $V$ .
- (c) A set of  $n$  elements forms a basis if and only if it spans  $V$ .
- (d) A set of  $n$  elements forms a basis if and only if it is linearly independent.

In other words, once we know the dimension of a vector space, to check that a collection having the correct number of elements forms a basis, we only need establish one of the two defining properties: span or linear independence. Thus,  $n$  elements that span an  $n$ -dimensional vector space are automatically linearly independent and hence form a basis; conversely,  $n$  linearly independent elements of an  $n$ -dimensional vector space automatically span the space and so form a basis.

**Example 2.32.** The standard basis of the space  $\mathcal{P}^{(n)}$  of polynomials of degree  $\leq n$  is given by the  $n + 1$  monomials  $1, x, x^2, \dots, x^n$ . We conclude that the vector space  $\mathcal{P}^{(n)}$

has dimension  $n + 1$ . Any other basis of  $\mathcal{P}^{(n)}$  must contain precisely  $n + 1$  polynomials. But, not every collection of  $n + 1$  polynomials in  $\mathcal{P}^{(n)}$  is a basis — they must be linearly independent. We conclude that no set of  $n$  or fewer polynomials can span  $\mathcal{P}^{(n)}$ , while any collection of  $n + 2$  or more polynomials of degree  $\leq n$  is automatically linearly dependent.

By definition, every vector space of dimension  $1 \leq n < \infty$  has a basis. If a vector space  $V$  has no basis, it is either the trivial vector space  $V = \{\mathbf{0}\}$ , which by convention has dimension 0, or its dimension is infinite. An infinite-dimensional vector space contains an infinite collection of linearly independent elements, and hence no (finite) basis. Examples of infinite-dimensional vector spaces include most spaces of functions, such as the spaces of continuous, differentiable, or mean zero functions, as well as the space of *all* polynomials, and the space of solutions to a linear homogeneous partial differential equation. (On the other hand, the solution space for a homogeneous linear ordinary differential equation turns out to be a finite-dimensional vector space.) There is a well-developed concept of a “complete basis” of certain infinite-dimensional function spaces, [67, 68], but this requires more delicate analytical considerations that lie beyond our present abilities. Thus, in this book, the term “basis” *always* means a finite collection of vectors in a finite-dimensional vector space.

**Proposition 2.33.** If  $\mathbf{v}_1, \dots, \mathbf{v}_m$  span the vector space  $V$ , then  $\dim V \leq m$ .

Thus, every vector space spanned by a finite number of elements is necessarily finite-dimensional, and so, if non-zero, admits a basis. Indeed, one can find the basis by successively looking at the members of a collection of spanning vectors, and retaining those that cannot be expressed as linear combinations of their predecessors in the list. Therefore,  $n = \dim V$  is the maximal number of linearly independent vectors in the set  $\mathbf{v}_1, \dots, \mathbf{v}_m$ . The details of the proof are left to the reader; see Exercise 2.4.22.

**Lemma 2.34.** The elements  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis of  $V$  if and only if every  $\mathbf{x} \in V$  can be written *uniquely* as a linear combination of the basis elements:

$$\mathbf{x} = c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n = \sum_{i=1}^n c_i \mathbf{v}_i. \quad (2.20)$$

*Proof:* The fact that a basis spans  $V$  implies that every  $\mathbf{x} \in V$  can be written as some linear combination of the basis elements. Suppose we can write an element

$$\mathbf{x} = c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n = \hat{c}_1 \mathbf{v}_1 + \dots + \hat{c}_n \mathbf{v}_n \quad (2.21)$$

as two different combinations. Subtracting one from the other, we obtain

$$(c_1 - \hat{c}_1) \mathbf{v}_1 + \dots + (c_n - \hat{c}_n) \mathbf{v}_n = \mathbf{0}.$$

The left-hand side is a linear combination of the basis elements, and hence vanishes if and only if all its coefficients  $c_i - \hat{c}_i = 0$ , meaning that the two linear combinations (2.21) are one and the same. *Q.E.D.*

One sometimes refers to the coefficients  $(c_1, \dots, c_n)$  in (2.20) as the *coordinates* of the vector  $\mathbf{x}$  with respect to the given basis. For the standard basis (2.17) of  $\mathbb{R}^n$ , the coordinates of a vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  are its entries, i.e., its usual Cartesian coordinates, cf. (2.18).

**Example 2.35.** *A Wavelet Basis.* The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad (2.22)$$

form a basis of  $\mathbb{R}^4$ . This is verified by performing Gaussian Elimination on the corresponding  $4 \times 4$  matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

to check that it is nonsingular. This is a very simple example of a *wavelet basis*. Wavelets play an increasingly central role in modern signal and digital image processing; see Section 9.7 and [18, 88].

How do we find the coordinates of a vector, say  $\mathbf{x} = (4, -2, 1, 5)^T$ , relative to the wavelet basis? We need to find the coefficients  $c_1, c_2, c_3, c_4$  such that

$$\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 + c_4 \mathbf{v}_4.$$

We use (2.13) to rewrite this equation in matrix form  $\mathbf{x} = A \mathbf{c}$ , where  $\mathbf{c} = (c_1, c_2, c_3, c_4)^T$ . Solving the resulting linear system by Gaussian Elimination produces

$$c_1 = 2, \quad c_2 = -1, \quad c_3 = 3, \quad c_4 = -2,$$

which are the coordinates of

$$\mathbf{x} = \begin{pmatrix} 4 \\ -2 \\ 1 \\ 5 \end{pmatrix} = 2\mathbf{v}_1 - \mathbf{v}_2 + 3\mathbf{v}_3 - 2\mathbf{v}_4 = 2 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + 3 \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} - 2 \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

in the wavelet basis. See Section 9.7 for the general theory of wavelet bases.

In general, to find the coordinates of a vector  $\mathbf{x}$  with respect to a new basis of  $\mathbb{R}^n$  requires the solution of a linear system of equations, namely

$$A \mathbf{c} = \mathbf{x} \quad \text{for} \quad \mathbf{c} = A^{-1} \mathbf{x}. \quad (2.23)$$

The columns of  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  are the basis vectors,  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  are the Cartesian coordinates of  $\mathbf{x}$ , with respect to the standard basis  $\mathbf{e}_1, \dots, \mathbf{e}_n$ , while  $\mathbf{c} = (c_1, c_2, \dots, c_n)^T$  contains its coordinates with respect to the new basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ . In practice, one finds the coordinates  $\mathbf{c}$  by Gaussian Elimination, *not* matrix inversion.

Why would one want to change bases? The answer is *simplification* and *speed* — many computations and formulas become much easier, and hence faster, to perform in a basis that is adapted to the problem at hand. In signal processing, wavelet bases are particularly appropriate for denoising, compression, and efficient storage of signals, including audio, still images, videos, medical and geophysical images, and so on. These processes would be quite time-consuming — if not impossible in complicated situations like video and three-dimensional image processing — to accomplish in the standard basis. Additional examples will appear throughout the text.

## Exercises

2.4.1. Determine which of the following sets of vectors are bases of  $\mathbb{R}^2$ : (a)  $\begin{pmatrix} 1 \\ -3 \end{pmatrix}, \begin{pmatrix} -2 \\ 5 \end{pmatrix}$ ;

(b)  $\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ ; (c)  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ ; (d)  $\begin{pmatrix} 3 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ ; (e)  $\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ .

2.4.2. Determine which of the following are bases of  $\mathbb{R}^3$ : (a)  $\begin{pmatrix} 2 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -5 \end{pmatrix}$ ,

(b)  $\begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 0 \end{pmatrix}$ ; (c)  $\begin{pmatrix} 0 \\ 4 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -8 \\ 1 \end{pmatrix}$ ; (d)  $\begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$ .

2.4.3. Let  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$ ,  $\mathbf{v}_4 = \begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix}$ . (a) Do  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$  span

$\mathbb{R}^3$ ? Why or why not? (b) Are  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$  linearly independent? Why or why not?

(c) Do  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$  form a basis for  $\mathbb{R}^3$ ? Why or why not? If not, is it possible to choose some subset that is a basis? (d) What is the dimension of the span of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ ? Justify your answer.

2.4.4. Answer Exercise 2.4.3 when  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \\ 5 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_4 = \begin{pmatrix} 1 \\ 3 \\ -1 \end{pmatrix}$ .

2.4.5. Find a basis for (a) the plane given by the equation  $z - 2y = 0$  in  $\mathbb{R}^3$ ; (b) the plane given by the equation  $4x + 3y - z = 0$  in  $\mathbb{R}^3$ ; (c) the hyperplane  $x + 2y + z - w = 0$  in  $\mathbb{R}^4$ .

2.4.6. (a) Show that  $\begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ , and  $\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}$  are two different bases for the plane

$x - 2y - 4z = 0$ . (b) Show how to write both elements of the second basis as linear combinations of the first. (c) Can you find a third basis?

2.4.7. A basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $\mathbb{R}^n$  is called *right-handed* if the  $n \times n$  matrix  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  whose columns are the basis vectors has positive determinant:  $\det A > 0$ . If  $\det A < 0$ , the basis is called *left-handed*. (a) Which of the following form right-handed bases of  $\mathbb{R}^3$ ?

(i)  $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$ , (iii)  $\begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$ ,

(iv)  $\begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$ . (b) Show that if  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  is a left-handed basis of  $\mathbb{R}^3$ , then  $\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_3$  and  $-\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  are both right-handed bases. (c) What sort of basis has  $\det A = 0$ ?

2.4.8. Find a basis for and the dimension of the following subspaces: (a) The space of solutions

to the linear system  $A\mathbf{x} = \mathbf{0}$ , where  $A = \begin{pmatrix} 1 & 2 & -1 & 1 \\ 3 & 0 & 2 & -1 \end{pmatrix}$ . (b) The set of all quadratic polynomials  $p(x) = ax^2 + bx + c$  that satisfy  $p(1) = 0$ . (c) The space of all solutions to the homogeneous ordinary differential equation  $u''' - u'' + 4u' - 4u = 0$ .

2.4.9. (a) Prove that  $1 + t^2, t + t^2, 1 + 2t + t^2$  is a basis for the space of quadratic polynomials  $\mathcal{P}^{(2)}$ . (b) Find the coordinates of  $p(t) = 1 + 4t + 7t^2$  in this basis.

2.4.10. Find a basis for and the dimension of the span of

$$(a) \begin{pmatrix} 3 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -6 \\ -2 \\ 2 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ 1 \\ 1 \end{pmatrix}.$$

2.4.11. (a) Show that  $1, 1 - t, (1 - t)^2, (1 - t)^3$  is a basis for  $\mathcal{P}^{(3)}$ .

(b) Write  $p(t) = 1 + t^3$  in terms of the basis elements.

2.4.12. Let  $\mathcal{P}^{(4)}$  denote the vector space consisting of all polynomials  $p(x)$  of degree  $\leq 4$ .

(a) Are  $x^3 - 3x + 1, x^4 - 6x + 3, x^4 - 2x^3 + 1$  linearly independent elements of  $\mathcal{P}^{(4)}$ ?

(b) What is the dimension of the subspace of  $\mathcal{P}^{(4)}$  they span?

2.4.13. Let  $S = \left\{ 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4} \right\}$ . (a) Show that the sample vectors corresponding to the functions

$1, \cos \pi x, \cos 2\pi x$ , and  $\cos 3\pi x$  form a basis for the vector space of all sample functions on  $S$ . (b) Write the sampled version of the function  $f(x) = x$  in terms of this basis.

2.4.14. (a) Prove that the vector space of all  $2 \times 2$  matrices is a four-dimensional vector space by exhibiting a basis. (b) Generalize your result and prove that the vector space  $\mathcal{M}_{m \times n}$  consisting of all  $m \times n$  matrices has dimension  $mn$ .

2.4.15. Determine all values of the scalar  $k$  for which the following four matrices form a basis

$$\text{for } \mathcal{M}_{2 \times 2}: A_1 = \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} k & -3 \\ 1 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0 \\ -k & 2 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & k \\ -1 & -2 \end{pmatrix}.$$

2.4.16. Prove that the space of diagonal  $n \times n$  matrices is an  $n$ -dimensional vector space.

2.4.17. (a) Find a basis for and the dimension of the space of upper triangular  $2 \times 2$  matrices.

(b) Can you generalize your result to upper triangular  $n \times n$  matrices?

2.4.18. (a) What is the dimension of the vector space of  $2 \times 2$  symmetric matrices? Of skew-symmetric matrices? (b) Generalize to the  $3 \times 3$  case. (c) What about  $n \times n$  matrices?

◇ 2.4.19. A matrix is said to be a *semi-magic square* if its row sums and column sums (i.e., the sum of entries in an individual row or column) all add up to the same number. An example

$$\text{is } \begin{pmatrix} 8 & 1 & 6 \\ 3 & 5 & 7 \\ 4 & 9 & 2 \end{pmatrix}, \text{ whose row and column sums are all equal to 15. (a) Explain why the set}$$

of all semi-magic squares is a subspace of the vector space of  $3 \times 3$  matrices. (b) Prove that the  $3 \times 3$  permutation matrices (1.30) span the space of semi-magic squares. What is its dimension? (c) A *magic square* also has the diagonal and *anti-diagonal* (running from top right to bottom left) add up to the common row and column sum; the preceding  $3 \times 3$  example is magic. Does the set of  $3 \times 3$  magic squares form a vector space? If so, what is its dimension? (d) Write down a formula for all  $3 \times 3$  magic squares.

◇ 2.4.20. (a) Prove that if  $\mathbf{v}_1, \dots, \mathbf{v}_m$  forms a basis for  $V \subsetneq \mathbb{R}^n$ , then  $m < n$ . (b) Under the hypothesis of part (a), prove that there exist vectors  $\mathbf{v}_{m+1}, \dots, \mathbf{v}_n \in \mathbb{R}^n \setminus V$  such that the complete collection  $\mathbf{v}_1, \dots, \mathbf{v}_n$  forms a basis for  $\mathbb{R}^n$ . (c) Illustrate by constructing bases of  $\mathbb{R}^3$  that include (i) the basis  $(1, 1, \frac{1}{2})^T$  of the line  $x = y = 2z$ ; (ii) the basis  $(1, 0, -1)^T, (0, 1, -2)^T$  of the plane  $x + 2y + z = 0$ .

◇ 2.4.21. Suppose that  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis for  $\mathbb{R}^n$ . Let  $A$  be a nonsingular matrix. Prove that  $A\mathbf{v}_1, \dots, A\mathbf{v}_n$  also form a basis for  $\mathbb{R}^n$ . What is this basis if you start with the standard basis:  $\mathbf{v}_i = \mathbf{e}_i$ ?

◇ 2.4.22. Show that if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  span  $V \neq \{\mathbf{0}\}$ , then one can choose a subset  $\mathbf{v}_{i_1}, \dots, \mathbf{v}_{i_m}$  that forms a basis of  $V$ . Thus,  $\dim V = m \leq n$ . Under what conditions is  $\dim V = n$ ?

◇ 2.4.23. Prove that if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are a basis of  $V$ , then every subset thereof, e.g.,  $\mathbf{v}_{i_1}, \dots, \mathbf{v}_{i_k}$ , is linearly independent.

- ◊ 2.4.24. Show, by example, how the uniqueness result in Lemma 2.34 fails if one has a linearly dependent set of vectors.
- ◊ 2.4.25. Let  $W \subset V$  be a subspace. (a) Prove that  $\dim W \leq \dim V$ .  
 (b) Prove that if  $\dim W = \dim V = n < \infty$ , then  $W = V$ . Equivalently, if  $W \subsetneq V$  is a proper subspace of a finite-dimensional vector space, then  $\dim W < \dim V$ .  
 (c) Give an example in which the result is false if  $\dim V = \infty$ .
- ◊ 2.4.26. Let  $W, Z \subset V$  be complementary subspaces in a finite-dimensional vector space  $V$ , as in Exercise 2.2.24. (a) Prove that if  $\mathbf{w}_1, \dots, \mathbf{w}_j$  form a basis for  $W$  and  $\mathbf{z}_1, \dots, \mathbf{z}_k$  a basis for  $Z$ , then  $\mathbf{w}_1, \dots, \mathbf{w}_j, \mathbf{z}_1, \dots, \mathbf{z}_k$  form a basis for  $V$ . (b) Prove that  $\dim W + \dim Z = \dim V$ .
- ◊ 2.4.27. Let  $V$  be a finite-dimensional vector space and  $W \subset V$  a subspace. Prove that the quotient space, as defined in Exercise 2.2.29, has dimension  $\dim(V/W) = \dim V - \dim W$ .
- ◊ 2.4.28. Let  $f_1(x), \dots, f_n(x)$  be scalar functions. Suppose that *every* set of sample points  $x_1, \dots, x_m \in \mathbb{R}$ , for all finite  $m \geq 1$ , leads to linearly dependent sample vectors  $\mathbf{f}_1, \dots, \mathbf{f}_n \in \mathbb{R}^m$ . Prove that  $f_1(x), \dots, f_n(x)$  are linearly dependent functions.  
*Hint:* Given sample points  $x_1, \dots, x_m$ , let  $V_{x_1, \dots, x_m} \subset \mathbb{R}^n$  be the subspace consisting of all vectors  $\mathbf{c} = (c_1, c_2, \dots, c_n)^T$  such that  $c_1 \mathbf{f}_1 + \dots + c_n \mathbf{f}_n = \mathbf{0}$ . First, show that one can select sample points  $x_1, x_2, x_3, \dots$  such that  $\mathbb{R}^n \supsetneq V_{x_1} \supsetneq V_{x_1, x_2} \supsetneq \dots$ . Then, apply Exercise 2.4.25 to conclude that  $V_{x_1, \dots, x_n} = \{\mathbf{0}\}$ .

## 2.5 The Fundamental Matrix Subspaces

Let us now return to the general study of linear systems of equations, which we write in our usual matrix form

$$A \mathbf{x} = \mathbf{b}. \quad (2.24)$$

As before,  $A$  is an  $m \times n$  matrix, where  $m$  is the number of equations, so  $\mathbf{b} \in \mathbb{R}^m$ , and  $n$  is the number of unknowns, i.e., the entries of  $\mathbf{x} \in \mathbb{R}^n$ . We already know how to solve the system, at least when the coefficient matrix is not too large: just apply a variant of Gaussian Elimination. Our goal now is to better understand the solution(s) and thereby prepare ourselves for more sophisticated problems and solution techniques.

### Kernel and Image

There are four important vector subspaces associated with any matrix. The first two are defined as follows.

**Definition 2.36.** The *image* of an  $m \times n$  matrix  $A$  is the subspace  $\text{img } A \subset \mathbb{R}^m$  spanned by its columns. The *kernel* of  $A$  is the subspace  $\ker A \subset \mathbb{R}^n$  consisting of all vectors that are annihilated by  $A$ , so

$$\ker A = \{ \mathbf{z} \in \mathbb{R}^n \mid A \mathbf{z} = \mathbf{0} \} \subset \mathbb{R}^n. \quad (2.25)$$

The image is also known as the *column space* or the *range*<sup>†</sup> of the matrix. By definition,

<sup>†</sup> The latter term can be confusing, since some authors call all of  $\mathbb{R}^m$  the range of the (function defined by the) matrix, hence our preference to use image here, and, later, *codomain* to refer to the space  $\mathbb{R}^n$ . On the other hand, the space  $\mathbb{R}^m$  will be called the *domain* of the (function defined by the) matrix.

a vector  $\mathbf{b} \in \mathbb{R}^m$  belongs to  $\text{img } A$  if it can be written as a linear combination,

$$\mathbf{b} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n,$$

of the columns of  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ . By our basic matrix multiplication formula (2.13), the right-hand side of this equation equals the product  $A \mathbf{x}$  of the matrix  $A$  with the column vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ , and hence  $\mathbf{b} = A \mathbf{x}$  for some  $\mathbf{x} \in \mathbb{R}^n$ . Thus,

$$\text{img } A = \{ A \mathbf{x} \mid \mathbf{x} \in \mathbb{R}^n \} \subset \mathbb{R}^m, \quad (2.26)$$

and so *a vector  $\mathbf{b}$  lies in the image of  $A$  if and only if the linear system  $A \mathbf{x} = \mathbf{b}$  has a solution*. The compatibility conditions for linear systems can thereby be re-interpreted as the requirements for a vector to lie in the image of the coefficient matrix.

A common alternative name for the kernel is the *null space*. The kernel or null space of  $A$  is the set of solutions  $\mathbf{z}$  to the homogeneous system  $A \mathbf{z} = \mathbf{0}$ . The proof that  $\ker A$  is a subspace requires us to verify the usual closure conditions: suppose that  $\mathbf{z}, \mathbf{w} \in \ker A$ , so that  $A \mathbf{z} = \mathbf{0} = A \mathbf{w}$ . Then, by the compatibility of scalar and matrix multiplication, for any scalars  $c, d$ ,

$$A(c \mathbf{z} + d \mathbf{w}) = c A \mathbf{z} + d A \mathbf{w} = \mathbf{0},$$

which implies that  $c \mathbf{z} + d \mathbf{w} \in \ker A$ . Closure of  $\ker A$  can be re-expressed as the following important *superposition principle* for solutions to a homogeneous system of linear equations.

**Theorem 2.37.** If  $\mathbf{z}_1, \dots, \mathbf{z}_k$  are individual solutions to the *same* homogeneous linear system  $A \mathbf{z} = \mathbf{0}$ , then so is every linear combination  $c_1 \mathbf{z}_1 + \cdots + c_k \mathbf{z}_k$ .

**Warning.** The set of solutions to an inhomogeneous linear system  $A \mathbf{x} = \mathbf{b}$  with  $\mathbf{b} \neq \mathbf{0}$  is *not* a subspace. Linear combinations of solutions are not, in general, solutions to the same inhomogeneous system.

Superposition is the reason why linear systems are so much easier to solve, since one needs to find only relatively few solutions in order to construct the general solution as a linear combination. In Chapter 7 we shall see that superposition applies to completely general linear systems, including linear differential equations, both ordinary and partial; linear boundary value problems; linear integral equations; linear control systems; etc.

**Example 2.38.** Let us compute the kernel of the matrix

$$A = \begin{pmatrix} 1 & -2 & 0 & 3 \\ 2 & -3 & -1 & -4 \\ 3 & -5 & -1 & -1 \end{pmatrix}.$$

Our task is to solve the homogeneous system  $A \mathbf{x} = \mathbf{0}$ , so we need only perform the elementary row operations on  $A$  itself. The resulting row echelon form

$$U = \begin{pmatrix} 1 & -2 & 0 & 3 \\ 0 & 1 & -1 & -10 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

corresponds to the equations  $x - 2y + 3w = 0$ ,  $y - z - 10w = 0$ . The free variables are  $z, w$ , and the general solution is

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 2z + 17w \\ z + 10w \\ z \\ w \end{pmatrix} = z \begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \end{pmatrix} + w \begin{pmatrix} 17 \\ 10 \\ 0 \\ 1 \end{pmatrix}.$$

The result describes the most general vector in  $\ker A$ , which is thus the two-dimensional subspace of  $\mathbb{R}^4$  spanned by the linearly independent vectors  $(2, 1, 1, 0)^T$ ,  $(17, 10, 0, 1)^T$ . This example is indicative of a general method for finding a basis for  $\ker A$ , to be developed in more detail below.

Once we know the kernel of the coefficient matrix  $A$ , i.e., the space of solutions to the homogeneous system  $A\mathbf{z} = \mathbf{0}$ , we are able to completely characterize the solutions to the inhomogeneous linear system (2.24).

**Theorem 2.39.** The linear system  $A\mathbf{x} = \mathbf{b}$  has a solution  $\mathbf{x}^*$  if and only if  $\mathbf{b}$  lies in the image of  $A$ . If this occurs, then  $\mathbf{x}$  is a solution to the linear system if and only if

$$\mathbf{x} = \mathbf{x}^* + \mathbf{z}, \quad (2.27)$$

where  $\mathbf{z} \in \ker A$  is an element of the kernel of the coefficient matrix.

*Proof:* We already demonstrated the first part of the theorem. If  $A\mathbf{x} = \mathbf{b} = A\mathbf{x}^*$  are any two solutions, then their difference  $\mathbf{z} = \mathbf{x} - \mathbf{x}^*$  satisfies

$$A\mathbf{z} = A(\mathbf{x} - \mathbf{x}^*) = A\mathbf{x} - A\mathbf{x}^* = \mathbf{b} - \mathbf{b} = \mathbf{0},$$

and hence  $\mathbf{z}$  is in the kernel of  $A$ . Therefore,  $\mathbf{x}$  and  $\mathbf{x}^*$  are related by formula (2.27), which proves the second part of the theorem. *Q.E.D.*

Therefore, to construct the most general solution to an inhomogeneous system, we need only know one *particular solution*  $\mathbf{x}^*$ , along with the general solution  $\mathbf{z} \in \ker A$  to the corresponding homogeneous system. This construction should remind the reader of the method for solving inhomogeneous linear ordinary differential equations. Indeed, both linear algebraic systems and linear ordinary differential equations are but two particular instances in the general theory of linear systems, to be developed in Chapter 7.

**Example 2.40.** Consider the system  $A\mathbf{x} = \mathbf{b}$ , where

$$A = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 1 & -1 \\ 1 & -2 & 3 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

where the right-hand side of the system will remain unspecified for the moment. Applying our usual Gaussian Elimination procedure to the augmented matrix

$$\left( \begin{array}{ccc|c} 1 & 0 & -1 & b_1 \\ -1 & 1 & -1 & b_2 \\ 1 & -2 & 3 & b_3 \end{array} \right) \text{ leads to the row echelon form } \left( \begin{array}{ccc|c} 1 & 0 & -1 & b_1 \\ 0 & 1 & -2 & b_2 \\ 0 & 0 & 0 & b_3 + 2b_2 + b_1 \end{array} \right).$$

Therefore, the system has a solution if and only if the compatibility condition

$$b_1 + 2b_2 + b_3 = 0 \quad (2.28)$$

holds. This equation serves to characterize the vectors  $\mathbf{b}$  that belong to the image of the matrix  $A$ , which is therefore a plane in  $\mathbb{R}^3$ .

To characterize the kernel of  $A$ , we take  $\mathbf{b} = \mathbf{0}$ , and solve the homogeneous system

$A\mathbf{z} = \mathbf{0}$ . The row echelon form corresponds to the reduced system

$$z_1 - z_3 = 0, \quad z_2 - 2z_3 = 0.$$

The free variable is  $z_3$ , and the equations are solved to give

$$z_1 = c, \quad z_2 = 2c, \quad z_3 = c,$$

where  $c$  is an arbitrary scalar. Thus, the general solution to the homogeneous system is  $\mathbf{z} = (c, 2c, c)^T = c(1, 2, 1)^T$ , and so the kernel is the line in the direction of the vector  $(1, 2, 1)^T$ .

If we take  $\mathbf{b} = (3, -2, 1)^T$  — which satisfies (2.28) and hence lies in the image of  $A$  — then the general solution to the inhomogeneous system  $A\mathbf{x} = \mathbf{b}$  is

$$x_1 = 3 + c, \quad x_2 = 1 + 2c, \quad x_3 = c,$$

where  $c$  is arbitrary. We can write the solution in the form (2.27), namely

$$\mathbf{x} = \begin{pmatrix} 3+c \\ 1+2c \\ c \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \mathbf{x}^* + \mathbf{z}, \quad (2.29)$$

where, as in (2.27),  $\mathbf{x}^* = (3, 1, 0)^T$  plays the role of the particular solution, while  $\mathbf{z} = c(1, 2, 1)^T$  is the general element of the kernel.

Finally, we remark that the particular solution is not uniquely defined — any individual solution to the system will serve the purpose. Thus, in this example, we could choose, for instance,  $\mathbf{x}^{**} = (-2, -9, -5)^T$  instead, corresponding to  $c = -5$  in the preceding formula (2.29). The general solution can be expressed in the alternative form

$$\mathbf{x} = \mathbf{x}^{**} + \mathbf{z} = \begin{pmatrix} -2 \\ -9 \\ -5 \end{pmatrix} + \tilde{c} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \text{where} \quad \mathbf{z} = \tilde{c} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \in \ker A,$$

which agrees with (2.29) when we identify  $\tilde{c} = c + 5$ .

We can characterize the situations in which the linear system has a unique solution in any of the following equivalent ways.

**Proposition 2.41.** If  $A$  is an  $m \times n$  matrix, then the following conditions are equivalent:

- (i)  $\ker A = \{\mathbf{0}\}$ , i.e., the homogeneous system  $A\mathbf{x} = \mathbf{0}$  has the unique solution  $\mathbf{x} = \mathbf{0}$ .
- (ii)  $\text{rank } A = n$ .
- (iii) The linear system  $A\mathbf{x} = \mathbf{b}$  has no free variables.
- (iv) The system  $A\mathbf{x} = \mathbf{b}$  has a unique solution for each  $\mathbf{b} \in \text{img } A$ .

Thus, while existence of a solution may depend upon the particularities of the right-hand side  $\mathbf{b}$ , uniqueness is universal: if for any one  $\mathbf{b}$ , e.g.,  $\mathbf{b} = \mathbf{0}$ , the system admits a unique solution, then all  $\mathbf{b} \in \text{img } A$  also admit unique solutions. Specializing even further to square matrices, we can now characterize invertible matrices by looking at either their kernels or their images.

**Proposition 2.42.** If  $A$  is a square  $n \times n$  matrix, then the following four conditions are equivalent: (i)  $A$  is nonsingular; (ii)  $\text{rank } A = n$ ; (iii)  $\ker A = \{\mathbf{0}\}$ ; (iv)  $\text{img } A = \mathbb{R}^n$ .

## Exercises

2.5.1. Characterize the image and kernel of the following matrices:

$$(a) \begin{pmatrix} 8 & -4 \\ -6 & 3 \end{pmatrix}, (b) \begin{pmatrix} 1 & -1 & 2 \\ -2 & 2 & -4 \end{pmatrix}, (c) \begin{pmatrix} 1 & 2 & 3 \\ -2 & 4 & 1 \\ 4 & 0 & 5 \end{pmatrix}, (d) \begin{pmatrix} 1 & -1 & 0 & 1 \\ -1 & 0 & 1 & -1 \\ 1 & -2 & 1 & 1 \\ 1 & 2 & -3 & 1 \end{pmatrix}.$$

2.5.2. For the following matrices, write the kernel as the span of a finite number of vectors.

Is the kernel a point, line, plane, or all of  $\mathbb{R}^3$ ? (a)  $(2 \ 1 \ 5)$ , (b)  $\begin{pmatrix} 1 & 2 & -1 \\ 3 & -2 & 0 \end{pmatrix}$ ,  
 (c)  $\begin{pmatrix} 2 & 6 & -4 \\ -1 & -3 & 2 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 2 & 5 \\ 0 & 4 & 8 \\ 1 & -6 & -11 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & -2 \\ 3 & -1 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & -2 & 3 \\ -3 & 6 & -9 \\ -2 & 4 & -6 \\ 3 & 0 & -1 \end{pmatrix}$ .

2.5.3. (a) Find the kernel and image of the coefficient matrix for the system  $x - 3y + 2z = a$ ,  $2x - 6y + 2w = b$ ,  $z - 3w = c$ . (b) Write down compatibility conditions on  $a, b, c$  for a solution to exist.

2.5.4. Suppose  $\mathbf{x}^* = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$  is a particular solution to the equation  $\begin{pmatrix} 1 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \mathbf{x} = \mathbf{b}$ .

(a) What is  $\mathbf{b}$ ? (b) Find the general solution.

2.5.5. Prove that the average of all the entries in each row of  $A$  is 0 if and only if  $(1, 1, \dots, 1)^T \in \ker A$ .

2.5.6. *True or false:* If  $A$  is a square matrix, then  $\ker A \cap \text{img } A = \{\mathbf{0}\}$ .

2.5.7. Write the general solution to the following linear systems in the form (2.27). Clearly identify the particular solution  $\mathbf{x}^*$  and the element  $\mathbf{z}$  of the kernel. (a)  $x - y + 3z = 1$ ,

$$(b) \begin{pmatrix} 1 & -2 & 0 \\ 2 & 3 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}, (c) \begin{pmatrix} 1 & -1 & 0 \\ 2 & 0 & -4 \\ 2 & -1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ -6 \\ -4 \end{pmatrix},$$

$$(d) \begin{pmatrix} 2 & -1 & 1 \\ 4 & -1 & 2 \\ 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, (e) \begin{pmatrix} 1 & -2 \\ 2 & -4 \\ -3 & 6 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \\ 3 \\ 1 \end{pmatrix},$$

$$(f) \begin{pmatrix} 1 & -3 & 2 & 0 \\ -1 & 5 & 1 & 1 \\ 2 & -8 & 1 & -1 \end{pmatrix} \begin{pmatrix} p \\ q \\ r \\ s \end{pmatrix} = \begin{pmatrix} 4 \\ -3 \\ 7 \end{pmatrix}, (g) \begin{pmatrix} 0 & -1 & 2 & -1 \\ 1 & -3 & 0 & 1 \\ -2 & 5 & 2 & -3 \\ 1 & 1 & -8 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -2 \\ -3 \\ 4 \\ 5 \end{pmatrix}.$$

2.5.8. Given  $a, r \neq 0$ , characterize the kernel and the image of the matrix

$$\begin{pmatrix} a & ar & \dots & ar^{n-1} \\ ar^n & ar^{n+1} & \dots & ar^{2n-1} \\ \vdots & \vdots & \ddots & \vdots \\ ar^{(n-1)n} & ar^{(n-1)n+1} & \dots & ar^{n^2-1} \end{pmatrix}. \quad \text{Hint: See Exercise 1.8.17.}$$

◇ 2.5.9. Let the square matrix  $P$  be idempotent, meaning that  $P^2 = P$ . (a) Prove that  $\mathbf{w} \in \text{img } P$  if and only if  $P\mathbf{w} = \mathbf{w}$ . (b) Show that  $\text{img } P$  and  $\ker P$  are complementary subspaces, as defined in Exercise 2.2.24, so every  $\mathbf{v} \in \mathbb{R}^n$  can be uniquely written as  $\mathbf{v} = \mathbf{w} + \mathbf{z}$  where  $\mathbf{w} \in \text{img } P$ ,  $\mathbf{z} \in \ker P$ .

◇ 2.5.10. Let  $A$  be an  $m \times n$  matrix. Suppose that  $C = \begin{pmatrix} A \\ B \end{pmatrix}$  is an  $(m+k) \times n$  matrix whose

first  $m$  rows are the same as those of  $A$ . Prove that  $\ker C \subseteq \ker A$ . Thus, appending more rows cannot increase the size of a matrix's kernel. Give an example in which  $\ker C \neq \ker A$ .

- ◇ 2.5.11. Let  $A$  be an  $m \times n$  matrix. Suppose that  $C = (A \ B)$  is an  $m \times (n + k)$  matrix whose first  $n$  columns are the same as those of  $A$ . Prove that  $\text{img } C \supseteq \text{img } A$ . Thus, appending more columns cannot decrease the size of a matrix's image. Give an example in which  $\text{img } C \neq \text{img } A$ .

## The Superposition Principle

The principle of superposition lies at the heart of linearity. For homogeneous systems, superposition allows one to generate new solutions by combining known solutions. For inhomogeneous systems, superposition combines the solutions corresponding to different inhomogeneities.

Suppose we know particular solutions  $\mathbf{x}_1^*$  and  $\mathbf{x}_2^*$  to two inhomogeneous linear systems

$$A \mathbf{x} = \mathbf{b}_1, \quad A \mathbf{x} = \mathbf{b}_2,$$

that have the *same* coefficient matrix  $A$ . Consider the system

$$A \mathbf{x} = c_1 \mathbf{b}_1 + c_2 \mathbf{b}_2,$$

whose right-hand side is a linear combination, or *superposition*, of the previous two. Then a particular solution to the combined system is given by the *same* superposition of the previous solutions:

$$\mathbf{x}^* = c_1 \mathbf{x}_1^* + c_2 \mathbf{x}_2^*.$$

The proof is easy:

$$A \mathbf{x}^* = A(c_1 \mathbf{x}_1^* + c_2 \mathbf{x}_2^*) = c_1 A \mathbf{x}_1^* + c_2 A \mathbf{x}_2^* = c_1 \mathbf{b}_1 + c_2 \mathbf{b}_2.$$

In physical applications, the inhomogeneities  $\mathbf{b}_1, \mathbf{b}_2$  typically represent external forces, and the solutions  $\mathbf{x}_1^*, \mathbf{x}_2^*$  represent the respective responses of the physical apparatus. The linear superposition principle says that if we know how the system responds to the individual forces, we immediately know its response to any combination thereof. The precise details of the system are irrelevant — all that is required is its linearity.

**Example 2.43.** For example, the system

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

models the mechanical response of a pair of masses connected by springs, subject to external forcing. The solution  $\mathbf{x} = (x_1, x_2)^T$  represents the displacements of the masses, while the entries of the right-hand side  $\mathbf{f} = (f_1, f_2)^T$  are the applied forces. (Details can be found in Chapter 6.) We can directly determine the response of the system  $\mathbf{x}_1^* = (\frac{4}{15}, -\frac{1}{15})^T$  to a unit force  $\mathbf{e}_1 = (1, 0)^T$  on the first mass, and the response  $\mathbf{x}_2^* = (-\frac{1}{15}, \frac{4}{15})^T$  to a unit force  $\mathbf{e}_2 = (0, 1)^T$  on the second mass. Superposition gives the response of the system to a general force, since we can write

$$\mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = f_1 \mathbf{e}_1 + f_2 \mathbf{e}_2 = f_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + f_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and hence

$$\mathbf{x} = f_1 \mathbf{x}_1^* + f_2 \mathbf{x}_2^* = f_1 \begin{pmatrix} \frac{4}{15} \\ -\frac{1}{15} \end{pmatrix} + f_2 \begin{pmatrix} -\frac{1}{15} \\ \frac{4}{15} \end{pmatrix} = \begin{pmatrix} \frac{4}{15} f_1 - \frac{1}{15} f_2 \\ -\frac{1}{15} f_1 + \frac{4}{15} f_2 \end{pmatrix}.$$

The preceding construction is easily extended to several inhomogeneities, and the result is the general *Superposition Principle* for inhomogeneous linear systems.

**Theorem 2.44.** Suppose that  $\mathbf{x}_1^*, \dots, \mathbf{x}_k^*$  are particular solutions to each of the inhomogeneous linear systems

$$A\mathbf{x} = \mathbf{b}_1, \quad A\mathbf{x} = \mathbf{b}_2, \quad \dots \quad A\mathbf{x} = \mathbf{b}_k, \quad (2.30)$$

all having the same coefficient matrix, and where  $\mathbf{b}_1, \dots, \mathbf{b}_k \in \text{img } A$ . Then, for any choice of scalars  $c_1, \dots, c_k$ , a particular solution to the combined system

$$A\mathbf{x} = c_1\mathbf{b}_1 + \dots + c_k\mathbf{b}_k \quad (2.31)$$

is the corresponding superposition

$$\mathbf{x}^* = c_1\mathbf{x}_1^* + \dots + c_k\mathbf{x}_k^* \quad (2.32)$$

of individual solutions. The general solution to (2.31) is

$$\mathbf{x} = \mathbf{x}^* + \mathbf{z} = c_1\mathbf{x}_1^* + \dots + c_k\mathbf{x}_k^* + \mathbf{z}, \quad (2.33)$$

where  $\mathbf{z} \in \ker A$  is the general solution to the homogeneous system  $A\mathbf{z} = \mathbf{0}$ .

For instance, if we know particular solutions  $\mathbf{x}_1^*, \dots, \mathbf{x}_m^*$  to

$$A\mathbf{x} = \mathbf{e}_i, \quad \text{for each } i = 1, \dots, m, \quad (2.34)$$

where  $\mathbf{e}_1, \dots, \mathbf{e}_m$  are the standard basis vectors of  $\mathbb{R}^m$ , then we can reconstruct a particular solution  $\mathbf{x}^*$  to the general linear system  $A\mathbf{x} = \mathbf{b}$  by first writing

$$\mathbf{b} = b_1\mathbf{e}_1 + \dots + b_m\mathbf{e}_m$$

as a linear combination of the basis vectors, and then using superposition to form

$$\mathbf{x}^* = b_1\mathbf{x}_1^* + \dots + b_m\mathbf{x}_m^*. \quad (2.35)$$

However, for linear algebraic systems, the practical value of this insight is rather limited. Indeed, in the case that  $A$  is square and nonsingular, the superposition formula (2.35) is merely a reformulation of the method of computing the inverse of the matrix. Indeed, the vectors  $\mathbf{x}_1^*, \dots, \mathbf{x}_m^*$  that satisfy (2.34) are just the columns of  $A^{-1}$  (why?), while (2.35) is precisely the solution formula  $\mathbf{x}^* = A^{-1}\mathbf{b}$  that we abandoned in practical computations, in favor of the more efficient Gaussian Elimination process. Nevertheless, this idea turns out to have important implications in more general situations, such as linear differential equations and boundary value problems.

## Exercises

- 2.5.12. Find the solution  $\mathbf{x}_1^*$  to the system  $\begin{pmatrix} 1 & 2 \\ -3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ , and the solution  $\mathbf{x}_2^*$  to  $\begin{pmatrix} 1 & 2 \\ -3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ . Express the solution to  $\begin{pmatrix} 1 & 2 \\ -3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$  as a linear combination of  $\mathbf{x}_1^*$  and  $\mathbf{x}_2^*$ .

2.5.13. Let  $A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 5 & -1 \\ 1 & 3 & 2 \end{pmatrix}$ . Given that  $\mathbf{x}_1^* = \begin{pmatrix} 5 \\ -1 \\ 2 \end{pmatrix}$  solves  $A\mathbf{x} = \mathbf{b}_1 = \begin{pmatrix} 1 \\ 3 \\ 6 \end{pmatrix}$  and

$\mathbf{x}_2^* = \begin{pmatrix} -11 \\ 5 \\ -1 \end{pmatrix}$  solves  $A\mathbf{x} = \mathbf{b}_2 = \begin{pmatrix} 0 \\ 4 \\ 2 \end{pmatrix}$ , find a solution to  $A\mathbf{x} = 2\mathbf{b}_1 + \mathbf{b}_2 = \begin{pmatrix} 2 \\ 10 \\ 14 \end{pmatrix}$ .

2.5.14. (a) Show that  $\mathbf{x}_1^* = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$  and  $\mathbf{x}_2^* = \begin{pmatrix} -3 \\ 3 \\ -2 \end{pmatrix}$  are particular solutions to the system

$$\begin{pmatrix} 2 & -1 & -5 \\ 1 & -4 & -6 \\ 3 & 2 & -4 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ -3 \\ 5 \end{pmatrix}. \quad (b) \text{ Find the general solution.}$$

2.5.15. A physical apparatus moves 2 meters under a force of 4 newtons. Assuming linearity, how far will it move under a force of 10 newtons?

2.5.16. Applying a unit external force in the horizontal direction moves a mass 3 units to the right, while applying a unit force in the vertical direction moves it up 2 units. Assuming linearity, where will the mass move under the applied force  $\mathbf{f} = (2, -3)^T$ ?

2.5.17. Suppose  $\mathbf{x}_1^*$  and  $\mathbf{x}_2^*$  are both solutions to  $A\mathbf{x} = \mathbf{b}$ . List all linear combinations of  $\mathbf{x}_1^*$  and  $\mathbf{x}_2^*$  that solve the system.

◇ 2.5.18. Let  $A$  be a nonsingular  $m \times m$  matrix. (a) Explain in detail why the solutions  $\mathbf{x}_1^*, \dots, \mathbf{x}_m^*$  to the systems (2.34) are the columns of the matrix inverse  $A^{-1}$ .

(b) Illustrate your argument in the case  $A = \begin{pmatrix} 0 & 1 & 2 \\ -1 & 1 & 3 \\ 1 & 0 & 1 \end{pmatrix}$ .

2.5.19. *True or false:* If  $\mathbf{x}_1^*$  solves  $A\mathbf{x} = \mathbf{c}$ , and  $\mathbf{x}_2^*$  solves  $B\mathbf{x} = \mathbf{d}$ , then  $\mathbf{x}^* = \mathbf{x}_1^* + \mathbf{x}_2^*$  solves  $(A + B)\mathbf{x} = \mathbf{c} + \mathbf{d}$ .

◇ 2.5.20. Under what conditions on the coefficient matrix  $A$  will the systems in (2.34) all have a solution?

## Adjoint Systems, Cokernel, and Coimage

A linear system of  $m$  equations in  $n$  unknowns is based on an  $m \times n$  coefficient matrix  $A$ . The transposed matrix  $A^T$  will be of size  $n \times m$ , and forms the coefficient matrix of an associated linear system, consisting of  $n$  equations in  $m$  unknowns.

**Definition 2.45.** The *adjoint*<sup>†</sup> to a linear system  $A\mathbf{x} = \mathbf{b}$  of  $m$  equations in  $n$  unknowns is the linear system

$$A^T \mathbf{y} = \mathbf{f} \tag{2.36}$$

consisting of  $n$  equations in  $m$  unknowns  $\mathbf{y} \in \mathbb{R}^m$  with right-hand side  $\mathbf{f} \in \mathbb{R}^n$ .

**Example 2.46.** Consider the linear system

$$\begin{aligned} x_1 - 3x_2 - 7x_3 + 9x_4 &= b_1, \\ x_2 + 5x_3 - 3x_4 &= b_2, \\ x_1 - 2x_2 - 2x_3 + 6x_4 &= b_3, \end{aligned} \tag{2.37}$$

<sup>†</sup> **Warning.** Some texts misuse the term “adjoint” to describe the *adjugate* or *cofactor matrix*, [80]. The constructions are completely unrelated, and the adjugate will play no role in this book.

of three equations in four unknowns. Its coefficient matrix

$$A = \begin{pmatrix} 1 & -3 & -7 & 9 \\ 0 & 1 & 5 & -3 \\ 1 & -2 & -2 & 6 \end{pmatrix} \quad \text{has transpose} \quad A^T = \begin{pmatrix} 1 & 0 & 1 \\ -3 & 1 & -2 \\ -7 & 5 & -2 \\ 9 & -3 & 6 \end{pmatrix}.$$

Thus, the adjoint system to (2.37) is the following system of four equations in three unknowns:

$$\begin{aligned} y_1 + y_3 &= f_1, \\ -3y_1 + y_2 - 2y_3 &= f_2, \\ -7y_1 + 5y_2 - 2y_3 &= f_3, \\ 9y_1 - 3y_2 + 6y_3 &= f_4. \end{aligned} \tag{2.38}$$

On the surface, there appears to be no direct connection between the solutions to a linear system and its adjoint. Nevertheless, as we shall soon see (and then in even greater depth in Sections 4.4 and 8.7), the two are linked in a number of remarkable, but subtle ways. As a first step in this direction, we use the adjoint system to define the remaining two fundamental subspaces associated with a coefficient matrix  $A$ .

**Definition 2.47.** The *coimage* of an  $m \times n$  matrix  $A$  is the image of its transpose,

$$\text{coimg } A = \text{img } A^T = \{ A^T \mathbf{y} \mid \mathbf{y} \in \mathbb{R}^m \} \subset \mathbb{R}^n. \tag{2.39}$$

The *cokernel* of  $A$  is the kernel of its transpose,

$$\text{coker } A = \ker A^T = \{ \mathbf{w} \in \mathbb{R}^m \mid A^T \mathbf{w} = \mathbf{0} \} \subset \mathbb{R}^m, \tag{2.40}$$

that is, the set of solutions to the homogeneous adjoint system.

The coimage coincides with the subspace of  $\mathbb{R}^n$  spanned by the rows<sup>†</sup> of  $A$ , and is thus often referred to as the *row space*. As a direct consequence of Theorem 2.39, the adjoint system  $A^T \mathbf{y} = \mathbf{f}$  has a solution if and only if  $\mathbf{f} \in \text{img } A^T = \text{coimg } A$ . The cokernel is also sometimes called the *left null space* of  $A$ , since it can be identified with the set of all row vectors  $\mathbf{r}$  satisfying  $\mathbf{r} A = \mathbf{0}^T$ , where  $\mathbf{0}^T$  is the row vector with  $m$  zero entries. Indeed, we can identify  $\mathbf{r} = \mathbf{w}^T$  and so, taking the transpose of the preceding equation, deduce  $A^T \mathbf{w} = (\mathbf{w}^T A)^T = (\mathbf{r} A)^T = \mathbf{0}$ , and so  $\mathbf{w} = \mathbf{r}^T \in \text{coker } A$ .

**Example 2.48.** To solve the linear system (2.37) just presented, we perform Gaussian

Elimination on the augmented matrix  $\left( \begin{array}{cccc|c} 1 & -3 & -7 & 9 & b_1 \\ 0 & 1 & 5 & -3 & b_2 \\ 1 & -2 & -2 & 6 & b_3 \end{array} \right)$ , reducing it to the row

echelon form  $\left( \begin{array}{cccc|c} 1 & -3 & -7 & 9 & b_1 \\ 0 & 1 & 5 & -3 & b_2 \\ 0 & 0 & 0 & 0 & b_3 - b_2 - b_1 \end{array} \right)$ . Thus, the system has a solution if and only if

$$-b_1 - b_2 + b_3 = 0,$$

<sup>†</sup> Or, more precisely, the column vectors obtained by transposing the rows.

which is required in order that  $\mathbf{b} \in \text{img } A$ . For such vectors, the general solution is

$$\mathbf{x} = \begin{pmatrix} b_1 + 3b_2 - 8x_3 \\ b_2 - 5x_3 + 3x_4 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 + 3b_2 \\ b_2 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -8 \\ -5 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 3 \\ 0 \\ 1 \end{pmatrix}.$$

In the second expression, the first vector represents a particular solution, while the two remaining terms constitute the general element of  $\ker A$ .

The solution to the adjoint system (2.38) is also obtained by Gaussian Elimination,

starting with its augmented matrix 
$$\left( \begin{array}{ccc|c} 1 & 0 & 1 & f_1 \\ -3 & 1 & -2 & f_2 \\ -7 & 5 & -2 & f_3 \\ 9 & -3 & 6 & f_4 \end{array} \right)$$
. The resulting row echelon form is 
$$\left( \begin{array}{ccc|c} 1 & 0 & 1 & f_1 \\ 0 & 1 & 1 & f_2 + 3f_1 \\ 0 & 0 & 0 & f_3 - 5f_2 - 8f_1 \\ 0 & 0 & 0 & f_4 + 3f_2 \end{array} \right)$$
. Thus, there are two consistency constraints required for a solution to the adjoint system:

$$-8f_1 - 5f_2 + f_3 = 0, \quad 3f_2 + f_4 = 0.$$

These are the conditions required for the right-hand side to belong to the coimage:  $\mathbf{f} \in \text{img } A^T = \text{coimg } A$ . If these conditions are satisfied, the adjoint system has the following general solution depending on the single free variable  $y_3$ :

$$\mathbf{y} = \begin{pmatrix} f_1 - y_3 \\ 3f_1 + f_2 - y_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ 3f_1 + f_2 \\ 0 \end{pmatrix} + y_3 \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}.$$

In the latter formula, the first term represents a particular solution, while the second is the general element of the cokernel  $\ker A^T = \text{coker } A$ .

## The Fundamental Theorem of Linear Algebra

The four fundamental subspaces associated with an  $m \times n$  matrix  $A$ , then, are its image, coimage, kernel, and cokernel. The image and cokernel are subspaces of  $\mathbb{R}^m$ , while the kernel and coimage are subspaces of  $\mathbb{R}^n$ . The *Fundamental Theorem of Linear Algebra*<sup>†</sup> states that their dimensions are determined by the rank (and size) of the matrix.

**Theorem 2.49.** Let  $A$  be an  $m \times n$  matrix, and let  $r$  be its rank. Then

$$\begin{aligned} \dim \text{coimg } A &= \dim \text{img } A = \text{rank } A = \text{rank } A^T = r, \\ \dim \ker A &= n - r, \quad \dim \text{coker } A = m - r. \end{aligned} \tag{2.41}$$

Thus, the rank of a matrix, i.e., the number of pivots, indicates the number of linearly independent columns, which, remarkably, is always the same as the number of linearly independent rows. A matrix and its transpose are guaranteed to have the same rank, i.e.,

---

<sup>†</sup> Not to be confused with the Fundamental Theorem of Algebra, which states that every (nonconstant) polynomial has a complex root; see [26].

the same number of pivots, despite the fact that their row echelon forms are quite different, and are almost never transposes of each other. Theorem 2.49 also establishes our earlier contention that the rank of a matrix is an *intrinsic* quantity, since it equals the common dimension of its image and coimage, and so does not depend on which specific elementary row operations are employed during the reduction process, nor on the final row echelon form.

Let us turn to the proof of the Fundamental Theorem 2.49. Since the dimension of a subspace is prescribed by the number of vectors in any basis, we need to relate bases of the fundamental subspaces to the rank of the matrix. Before trying to digest the general argument, it is better first to understand how to construct the required bases in a particular example. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 1 & 2 \\ -8 & 4 & -6 & -4 \\ 4 & -2 & 3 & 2 \end{pmatrix}. \quad \text{Its row echelon form } U = \begin{pmatrix} 2 & -1 & 1 & 2 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (2.42)$$

is obtained in the usual manner. There are two pivots, and thus the rank of  $A$  is  $r = 2$ .

*Kernel:* The general solution to the homogeneous system  $A\mathbf{x} = \mathbf{0}$  can be expressed as a linear combination of  $n - r$  linearly independent vectors, whose coefficients are the free variables for the system corresponding to the  $n - r$  columns without pivots. In fact, these vectors form a basis for the kernel, which thus has dimension  $n - r$ .

In our example, the pivots are in columns 1 and 3, and so the free variables are  $x_2, x_4$ . Applying Back Substitution to the reduced homogeneous system  $U\mathbf{x} = \mathbf{0}$ , we obtain the general solution

$$\mathbf{x} = \begin{pmatrix} \frac{1}{2}x_2 - 2x_4 \\ x_2 \\ 2x_4 \\ x_4 \end{pmatrix} = x_2 \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -2 \\ 0 \\ 2 \\ 1 \end{pmatrix} \quad (2.43)$$

written as a linear combination of the vectors

$$\mathbf{z}_1 = \left( \frac{1}{2}, 1, 0, 0 \right)^T, \quad \mathbf{z}_2 = (-2, 0, 2, 1)^T.$$

We claim that  $\mathbf{z}_1, \mathbf{z}_2$  form a basis of  $\ker A$ . By construction, they span the kernel, and linear independence follows easily, since the only way in which the linear combination (2.43) could vanish is if both free variables vanish:  $x_2 = x_4 = 0$ .

*Coimage:* The coimage is the subspace of  $\mathbb{R}^n$  spanned by the rows<sup>†</sup> of  $A$ . As we prove below, applying an elementary row operation to a matrix does not alter its coimage. Since the row echelon form  $U$  is obtained from  $A$  by a sequence of elementary row operations, we conclude that  $\text{coimg } A = \text{coimg } U$ . Moreover, the row echelon structure implies that the  $r$  nonzero rows of  $U$  are necessarily linearly independent, and hence form a basis of both  $\text{coimg } U$  and  $\text{coimg } A$ , which therefore have dimension  $r = \text{rank } A$ . In our example, then, a basis for  $\text{coimg } A$  consists of the vectors

$$\mathbf{s}_1 = (2, -1, 1, 2)^T, \quad \mathbf{s}_2 = (0, 0, -2, 4)^T,$$

---

<sup>†</sup> Or, more correctly, the transposes of the rows, since the elements of  $\mathbb{R}^n$  are supposed to be column vectors.

coming from the nonzero rows of  $U$ . The reader can easily check their linear independence, as well as the fact that every row of  $A$  lies in their span.

*Image:* There are two methods for computing a basis of the image, or column space. The first proves that it has dimension equal to the rank. This has the important, and remarkable consequence that the space spanned by the rows of a matrix and the space spanned by its columns always have the same dimension, even though they are usually different subspaces of different vector spaces.

Now, the row echelon structure implies that the columns of  $U$  that contain the pivots form a basis for its image, i.e.,  $\text{img } U$ . In our example, these are its first and third columns, and you can check that they are linearly independent and span the full column space. But the image of  $A$  is *not* the same as the image of  $U$ , and so, unlike the coimage, we cannot directly use a basis for  $\text{img } U$  as a basis for  $\text{img } A$ . However, the linear dependencies among the columns of  $A$  and  $U$  are the same, and this implies that the  $r$  columns of  $A$  that end up containing the pivots will form a basis for  $\text{img } A$ . In our example (2.42), the pivots lie in the first and third columns of  $U$ , and hence the first and third columns of  $A$ ; namely,

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -8 \\ 4 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -6 \\ 3 \end{pmatrix},$$

form a basis for  $\text{img } A$ . This means that every column of  $A$  can be written uniquely as a linear combination of its first and third columns. Again, skeptics may wish to check this.

An alternative method to find a basis for the image is to recall that  $\text{img } A = \text{coimg } A^T$ , and hence we can employ the previous algorithm to compute  $\text{coimg } A^T$ . In our example, applying Gaussian Elimination to

$$A^T = \begin{pmatrix} 2 & -8 & 4 \\ -1 & 4 & -2 \\ 1 & -6 & 3 \\ 2 & -4 & 2 \end{pmatrix} \quad \text{leads to the row echelon form} \quad \widehat{U} = \begin{pmatrix} 2 & -8 & 4 \\ 0 & -2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (2.44)$$

Note that the row echelon form of  $A^T$  is *not* the transpose of the row echelon form of  $A$ . However, they do have the same number of pivots, since, as we now know, both  $A$  and  $A^T$  have the same rank, namely 2. The two nonzero rows of  $\widehat{U}$  (again transposed to be column vectors) form a basis for  $\text{coimg } A^T$ , and therefore

$$\mathbf{y}_1 = \begin{pmatrix} 2 \\ -8 \\ 4 \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix},$$

forms an alternative basis for  $\text{img } A$ .

*Cokernel:* Finally, to determine a basis for the cokernel, we apply the algorithm for finding a basis for  $\ker A^T = \text{coker } A$ . Since the ranks of  $A$  and  $A^T$  coincide, there are now  $m - r$  free variables, which is the same as the dimension of  $\ker A^T$ . In our particular example, using the reduced form (2.44), the only free variable is  $y_3$ , and the general solution to the homogeneous adjoint system  $A^T \mathbf{y} = \mathbf{0}$  is

$$\mathbf{y} = \begin{pmatrix} 0 \\ \frac{1}{2} y_3 \\ y_3 \end{pmatrix} = y_3 \begin{pmatrix} 0 \\ \frac{1}{2} \\ 1 \end{pmatrix}.$$

We conclude that  $\text{coker } A$  is one-dimensional, with basis  $(0, \frac{1}{2}, 1)^T$ .

- Summarizing, given an  $m \times n$  matrix  $A$  with row echelon form  $U$ , to find a basis for
- $\text{img } A$ : choose the  $r$  columns of  $A$  in which the pivots appear in  $U$ ;
  - $\ker A$ : write the general solution to  $A\mathbf{x} = \mathbf{0}$  as a linear combination of the  $n - r$  basis vectors whose coefficients are the free variables;
  - $\text{coimg } A$ : choose the  $r$  nonzero rows of  $U$ ;
  - $\text{coker } A$ : write the general solution to the adjoint system  $A^T\mathbf{y} = \mathbf{0}$  as a linear combination of the  $m - r$  basis vectors whose coefficients are the free variables. (An alternative method — one that does not require solving the adjoint system — can be found on page 223.)

Let us conclude this section by justifying these constructions for general matrices, and thereby complete the proof of the Fundamental Theorem 2.49.

*Kernel:* If  $A$  has rank  $r$ , then the general element of the kernel, i.e., solution to the homogeneous system  $A\mathbf{x} = \mathbf{0}$ , can be written as a linear combination of  $n - r$  vectors whose coefficients are the free variables, and hence these vectors span  $\ker A$ . Moreover, the only combination that yields the zero solution  $\mathbf{x} = \mathbf{0}$  is when all the free variables are zero, since any nonzero value for a free variable, say  $x_i \neq 0$ , gives a solution  $\mathbf{x} \neq \mathbf{0}$  whose  $i^{\text{th}}$  entry (at least) is nonzero. Thus, the only linear combination of the  $n - r$  kernel basis vectors that sums to  $\mathbf{0}$  is the trivial one, which implies their linear independence.

*Coimage:* We need to prove that elementary row operations do not change the coimage. To see this for row operations of the first type, suppose, for instance, that  $\hat{A}$  is obtained by adding  $b$  times the first row of  $A$  to the second row. If  $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_m$  are the rows of  $A$ , then the rows of  $\hat{A}$  are  $\mathbf{r}_1, \hat{\mathbf{r}}_2 = \mathbf{r}_2 + b\mathbf{r}_1, \mathbf{r}_3, \dots, \mathbf{r}_m$ . If

$$\mathbf{v} = c_1 \mathbf{r}_1 + c_2 \mathbf{r}_2 + c_3 \mathbf{r}_3 + \dots + c_m \mathbf{r}_m$$

is any vector belonging to  $\text{coimg } A$ , then

$$\mathbf{v} = \hat{c}_1 \mathbf{r}_1 + c_2 \hat{\mathbf{r}}_2 + c_3 \mathbf{r}_3 + \dots + c_m \mathbf{r}_m, \quad \text{where} \quad \hat{c}_1 = c_1 - b c_2,$$

is also a linear combination of the rows of the new matrix, and hence lies in  $\text{coimg } \hat{A}$ . The converse is also valid —  $\mathbf{v} \in \text{coimg } \hat{A}$  implies  $\mathbf{v} \in \text{coimg } A$  — and we conclude that elementary row operations of type #1 do not change  $\text{coimg } A$ . The proofs for the other two types of elementary row operations are even easier, and are left to the reader.

The basis for  $\text{coimg } A$  will be the first  $r$  nonzero pivot rows  $\mathbf{s}_1, \dots, \mathbf{s}_r$  of  $U$ . Since the other rows, if any, are all  $\mathbf{0}$ , the pivot rows clearly span  $\text{coimg } U = \text{coimg } A$ . To prove their linear independence, suppose

$$c_1 \mathbf{s}_1 + \dots + c_r \mathbf{s}_r = \mathbf{0}. \quad (2.45)$$

Let  $u_{1k} \neq 0$  be the first pivot. Since all entries of  $U$  lying below the pivot are zero, the  $k^{\text{th}}$  entry of (2.45) is  $c_1 u_{1k} = 0$ , which implies that  $c_1 = 0$ . Next, suppose  $u_{2l} \neq 0$  is the second pivot. Again, using the row echelon structure of  $U$ , the  $l^{\text{th}}$  entry of (2.45) is found to be  $c_1 u_{1l} + c_2 u_{2l} = 0$ , and so  $c_2 = 0$ , since we already know  $c_1 = 0$ . Continuing in this manner, we deduce that only the trivial linear combination  $c_1 = \dots = c_r = 0$  will satisfy (2.45), proving linear independence. Thus,  $\mathbf{s}_1, \dots, \mathbf{s}_r$  form a basis for  $\text{coimg } U = \text{coimg } A$ , which therefore has dimension  $r = \text{rank } A$ .

*Image:* In general, a vector  $\mathbf{b} \in \text{img } A$  if and only if it can be written as a linear combination of the columns:  $\mathbf{b} = A\mathbf{x}$ . But, as we know, the general solution to the linear

system  $A\mathbf{x} = \mathbf{b}$  is expressed in terms of the free and basic variables; in particular, we are allowed to set all the free variables to zero, and so end up writing  $\mathbf{b}$  in terms of the basic variables alone. This effectively expresses  $\mathbf{b}$  as a linear combination of the pivot columns of  $A$  only, which proves that they span  $\text{img } A$ . To prove their linear independence, suppose some linear combination of the pivot columns adds up to  $\mathbf{0}$ . Interpreting the coefficients as basic variables, this would correspond to a vector  $\mathbf{x}$ , all of whose free variables are zero, satisfying  $A\mathbf{x} = \mathbf{0}$ . But our solution to this homogeneous system expresses the basic variables as combinations of the free variables, which, if the latter are all zero, are also zero when the right-hand sides all vanish. This shows that, under these assumptions,  $\mathbf{x} = \mathbf{0}$ , and hence the pivot columns are linearly independent.

*Cokernel:* By the preceding arguments,  $\text{rank } A = \text{rank } A^T = r$ , and hence the general element of  $\text{coker } A = \ker A^T$  can be written as a linear combination of  $m - r$  basis vectors whose coefficients are the free variables in the homogeneous adjoint system  $A^T\mathbf{y} = \mathbf{0}$ . Linear independence of the basis elements follows as in the case of the kernel.

## Exercises

2.5.21. For each of the following matrices find bases for the (i) image, (ii) coimage, (iii) kernel, and (iv) cokernel.

$$(a) \begin{pmatrix} 1 & -3 \\ 2 & -6 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 0 & -8 \\ 1 & 2 & -1 \\ 2 & 4 & 6 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 1 & 2 & 1 \\ 1 & 0 & -1 & 3 \\ 2 & 3 & 7 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & -3 & 2 & 2 & 1 \\ 0 & 3 & -6 & 0 & -2 \\ 2 & -3 & -2 & 4 & 0 \\ 3 & -3 & -6 & 6 & 3 \\ 1 & 0 & -4 & 2 & 3 \end{pmatrix}.$$

2.5.22. Find a set of columns of the matrix  $\begin{pmatrix} -1 & 2 & 0 & -3 & 5 \\ 2 & -4 & 1 & 1 & -4 \\ -3 & 6 & 2 & 0 & 8 \end{pmatrix}$  that form a basis for its image. Then express each column as a linear combination of the basis columns.

2.5.23. For each of the following matrices  $A$ : (a) Determine the rank and the dimensions of the four fundamental subspaces. (b) Find bases for both the kernel and cokernel. (c) Find explicit conditions on vectors  $\mathbf{b}$  that guarantee that the system  $A\mathbf{x} = \mathbf{b}$  has a solution. (d) Write down a specific *nonzero* vector  $\mathbf{b}$  that satisfies your conditions, and then find all possible solutions  $\mathbf{x}$ .

$$(i) \begin{pmatrix} 1 & 2 \\ -2 & -4 \end{pmatrix}, \quad (ii) \begin{pmatrix} 3 & -1 & -2 \\ -6 & 2 & 4 \end{pmatrix}, \quad (iii) \begin{pmatrix} 1 & 5 \\ -2 & 3 \\ 2 & 7 \end{pmatrix}, \quad (iv) \begin{pmatrix} 2 & -5 & -1 \\ 1 & -6 & -4 \\ 3 & -4 & 2 \end{pmatrix},$$

$$(v) \begin{pmatrix} 2 & 5 & 7 \\ 6 & 13 & 19 \\ 3 & 8 & 11 \\ 1 & 2 & 3 \end{pmatrix}, \quad (vi) \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \\ 1 & -2 & 2 & 7 \\ 3 & 6 & 5 & -2 \end{pmatrix}, \quad (vii) \begin{pmatrix} 2 & 4 & 0 & -6 & 0 \\ 1 & 2 & 3 & 15 & 0 \\ 3 & 6 & -1 & 15 & 5 \\ -3 & -6 & 2 & 21 & -6 \end{pmatrix}.$$

2.5.24. Find the dimension of and a basis for the subspace spanned by the following sets of vectors. *Hint:* First identify the subspace with the image of a certain matrix.

$$(a) \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ -3 \end{pmatrix},$$

$$(d) \begin{pmatrix} 1 \\ 0 \\ -3 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \\ -3 \end{pmatrix}, \begin{pmatrix} -3 \\ -4 \\ 1 \\ 6 \end{pmatrix}, \begin{pmatrix} 1 \\ -3 \\ -8 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ -6 \\ 9 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 \\ 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ -3 \\ 4 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \end{pmatrix}.$$

2.5.25. Show that the set of all vectors  $\mathbf{v} = (a - 3b, a + 2c + 4d, b + 3c - d, c - d)^T$ , where  $a, b, c, d$  are real numbers, forms a subspace of  $\mathbb{R}^4$ , and find its dimension.

2.5.26. Find a basis of the solution space of the following homogeneous linear systems.

$$(a) \begin{array}{l} x_1 - 2x_3 = 0, \\ x_2 + x_4 = 0. \end{array} \quad (b) \begin{array}{l} 2x_1 + x_2 - 3x_3 + x_4 = 0, \\ 2x_1 - x_2 - x_3 - x_4 = 0. \end{array} \quad (c) \begin{array}{l} x_1 - x_2 - 2x_3 + 4x_4 = 0, \\ 2x_1 + x_2 - x_4 = 0, \\ -2x_1 + 2x_3 - 2x_4 = 0. \end{array}$$

2.5.27. Find bases for the image and coimage of  $\begin{pmatrix} 1 & -3 & 0 \\ 2 & -6 & 4 \\ -3 & 9 & 1 \end{pmatrix}$ . Make sure they have the same number of elements. Then write each row and column as a linear combination of the appropriate basis vectors.

2.5.28. Find bases for the image of  $\begin{pmatrix} 1 & 2 & -1 \\ 0 & 3 & -3 \\ 2 & -4 & 6 \\ 1 & 5 & -4 \end{pmatrix}$  using both of the indicated methods.

Demonstrate that they are indeed both bases for the same subspace by showing how to write each basis in terms of the other.

2.5.29. Show that  $\mathbf{v}_1 = (1, 2, 0, -1)^T, \mathbf{v}_2 = (-3, 1, 1, -1)^T, \mathbf{v}_3 = (2, 0, -4, 3)^T$  and  $\mathbf{w}_1 = (3, 2, -4, 2)^T, \mathbf{w}_2 = (2, 3, -7, 4)^T, \mathbf{w}_3 = (0, 3, -3, 1)^T$  are two bases for the same three-dimensional subspace  $V \subset \mathbb{R}^4$ .

2.5.30. (a) Prove that if  $A$  is a symmetric matrix, then  $\ker A = \text{coker } A$  and  $\text{img } A = \text{coimg } A$ .  
 (b) Use this observation to produce bases for the four fundamental subspaces associated

with  $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 2 \end{pmatrix}$ . (c) Is the converse to part (a) true?

2.5.31. (a) Write down a matrix of rank  $r$  whose first  $r$  rows do *not* form a basis for its row space. (b) Can you find an example that can be reduced to row echelon form without any row interchanges?

2.5.32. Let  $A$  be a  $4 \times 4$  matrix and let  $U$  be its row echelon form. (a) Suppose columns 1, 2, 4 of  $U$  form a basis for its image. Do columns 1, 2, 4 of  $A$  form a basis for its image? If so, explain why; if not, construct a counterexample. (b) Suppose rows 1, 2, 3 of  $U$  form a basis for its coimage. Do rows 1, 2, 3 of  $A$  form a basis for its coimage? If so, explain why; if not, construct a counterexample. (c) Suppose you find a basis for  $\ker U$ . Is it also a basis for  $\ker A$ ? (d) Suppose you find a basis for  $\text{coker } U$ . Is it also a basis for  $\text{coker } A$ ?

2.5.33. Can you devise a nonzero matrix whose row echelon form is the same as the row echelon form of its transpose?

◇ 2.5.34. Explain why the elementary row operations of types #2 and #3 do not change the coimage of a matrix.

2.5.35. Let  $A$  be an  $m \times n$  matrix. Prove that  $\text{img } A = \mathbb{R}^m$  if and only if  $\text{rank } A = m$ .

2.5.36. Prove or give a counterexample: If  $U$  is the row echelon form of  $A$ , then  $\text{img } U = \text{img } A$ .

◇ 2.5.37. (a) Devise an alternative method for finding a basis of the coimage of a matrix.

*Hint:* Look at the two methods for finding a basis for the image. (b) Use your method to find a basis for the coimage of  $\begin{pmatrix} 1 & 3 & -5 & 2 \\ 2 & -1 & 1 & -4 \\ 4 & 5 & -9 & 2 \end{pmatrix}$ . Is it the same basis as found by the method in the text?

◇ 2.5.38. Prove that  $\ker A \subseteq \ker A^2$ . More generally, prove  $\ker A \subseteq \ker B A$  for every compatible matrix  $B$ .

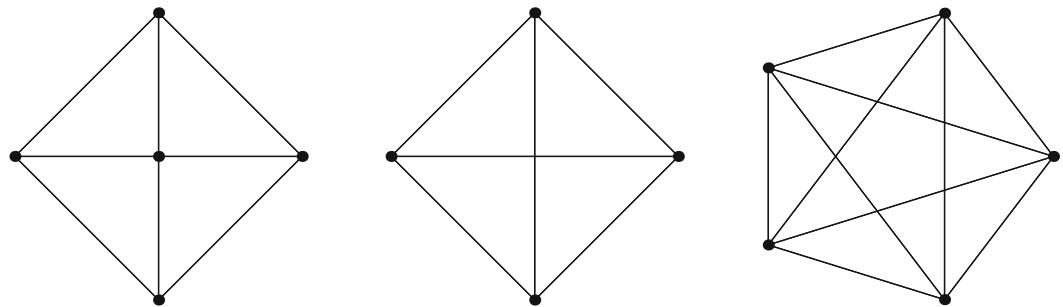
- ◇ 2.5.39. Prove that  $\text{img } A \supseteq \text{img } A^2$ . More generally, prove  $\text{img } A \supseteq \text{img } (AB)$  for every compatible matrix  $B$ .
- 2.5.40. Suppose  $A$  is an  $m \times n$  matrix, and  $B$  and  $C$  are nonsingular matrices of sizes  $m \times m$  and  $n \times n$ , respectively. Prove that  $\text{rank } A = \text{rank } BA = \text{rank } AC = \text{rank } BAC$ .
- 2.5.41. *True or false:* If  $\ker A = \ker B$ , then  $\text{rank } A = \text{rank } B$ .
- ◇ 2.5.42. Let  $A$  and  $B$  be matrices of respective sizes  $m \times n$  and  $n \times p$ .
- Prove that  $\dim \ker(AB) \leq \dim \ker A + \dim \ker B$ .
  - Prove the *Sylvester Inequalities*  $\text{rank } A + \text{rank } B - n \leq \text{rank}(AB) \leq \min\{\text{rank } A, \text{rank } B\}$ .
- ◇ 2.5.43. Suppose  $A$  is a nonsingular  $n \times n$  matrix. (a) Prove that every  $n \times (n+k)$  matrix of the form  $(A \ B)$ , where  $B$  has size  $n \times k$ , has rank  $n$ . (b) Prove that every  $(n+k) \times n$  matrix of the form  $\begin{pmatrix} A \\ C \end{pmatrix}$ , where  $C$  has size  $k \times n$ , has rank  $n$ .
- ◇ 2.5.44. Let  $A$  be an  $m \times n$  matrix of rank  $r$ . Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are a basis for  $\mathbb{R}^n$  such that  $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$  form a basis for  $\ker A$ . Prove that  $\mathbf{w}_1 = A\mathbf{v}_1, \dots, \mathbf{w}_r = A\mathbf{v}_r$  form a basis for  $\text{img } A$ .
- ◇ 2.5.45. (a) Suppose  $A, B$  are  $m \times n$  matrices such that  $\ker A = \ker B$ . Prove that there is a nonsingular  $m \times m$  matrix  $M$  such that  $MA = B$ . *Hint:* Use Exercise 2.5.44. (b) Use this to conclude that if  $A\mathbf{x} = \mathbf{b}$  and  $B\mathbf{x} = \mathbf{c}$  have the same solutions then they are equivalent linear systems, i.e., one can be obtained from the other by a sequence of elementary row operations.
- ◇ 2.5.46. (a) Let  $A$  be an  $m \times n$  matrix and let  $V$  be a subspace of  $\mathbb{R}^n$ . Show that  $W = AV = \{A\mathbf{v} \mid \mathbf{v} \in V\}$  forms a subspace of  $\text{img } A$ . (b) If  $\dim V = k$ , show that  $\dim W \leq \min\{k, r\}$ , where  $r = \text{rank } A$ . Give an example in which  $\dim(AV) < \dim V$ . *Hint:* Use Exercise 2.4.25.
- ◇ 2.5.47. (a) Show that an  $m \times n$  matrix has a left inverse if and only if it has rank  $n$ .  
*Hint:* Use Exercise 2.5.46. (b) Show that it has a right inverse if and only if it has rank  $m$ .  
(c) Conclude that only nonsingular square matrices have both left and right inverses.

## 2.6 Graphs and Digraphs

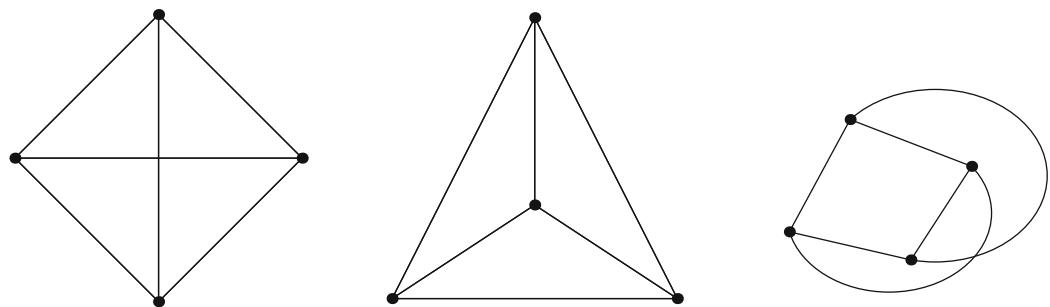
We now present an intriguing application of linear algebra to graph theory. A *graph* consists of a finite number of points, called *vertices*, and finitely many lines or curves connecting them, called *edges*. Each edge connects exactly two vertices, which are its endpoints. To avoid technicalities, we will always assume that the graph is *simple*, which means that every edge connects two *distinct* vertices, so no edge forms a *loop* that connects a vertex to itself, and, moreover, two distinct vertices are connected by at most one edge. Some examples of graphs appear in [Figure 2.6](#); the vertices are the black dots and the edges are the lines connecting them.

Graphs arise in a multitude of applications. A particular case that will be considered in depth is electrical networks, where the edges represent wires, and the vertices represent the nodes where the wires are connected. Another example is the framework for a building — the edges represent the beams, and the vertices the joints where the beams are connected. In each case, the graph encodes the topology — meaning interconnectedness — of the system, but not its geometry — lengths of edges, angles, etc.

In a planar representation of a graph, the edges are allowed to cross over each other at non-nodal points without meeting — think of a network where the (insulated) wires lie



**Figure 2.6.** Three Different Graphs.



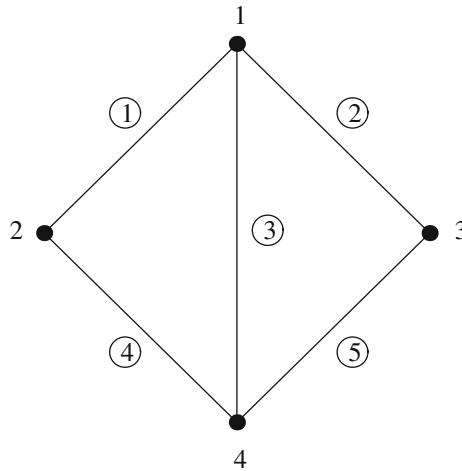
**Figure 2.7.** Three Versions of the Same Graph.

on top of each other, but do not interconnect. Thus, the first graph in Figure 2.6 has 5 vertices and 8 edges; the second has 4 vertices and 6 edges — the two central edges do not meet; the final graph has 5 vertices and 10 edges.

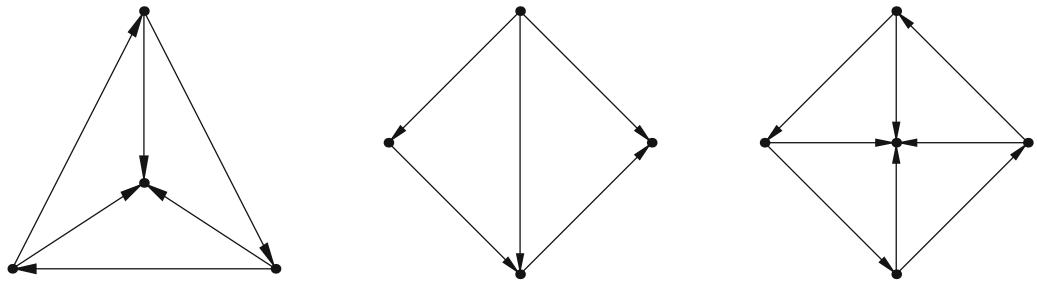
Two graphs are considered to be the same if there is a one-to-one correspondence between their edges and their vertices, so that matched edges connect matched vertices. In an electrical network, moving the nodes and wires around without cutting or rejoining will have no effect on the underlying graph. Consequently, there are many ways to draw a given graph; three representations of one and the same graph appear in Figure 2.7.

A *path* in a graph is an ordered list of distinct edges  $e_1, \dots, e_k$  connecting (not necessarily distinct) vertices  $v_1, \dots, v_{k+1}$  so that edge  $e_i$  connects vertex  $v_i$  to  $v_{i+1}$ . For instance, in the graph in Figure 2.8, one path starts at vertex 1, then goes in order along the edges labeled as 1, 4, 3, 2, successively passing through the vertices 1, 2, 4, 1, 3. Observe that while an edge cannot be repeated in a path, a vertex may be. A graph is *connected* if you can get from any vertex to any other vertex by a path, which is the most important case for applications. We note that every graph can be decomposed into a disconnected collection of connected subgraphs.

A *circuit* is a path that ends up where it began, i.e.,  $v_{k+1} = v_1$ . For example, the circuit in Figure 2.8 consisting of edges 1, 4, 5, 2 starts at vertex 1, then goes to vertices 2, 4, 3 in order, and finally returns to vertex 1. In a closed circuit, the choice of starting vertex is not important, and we identify circuits that go around the edges in the same order. Thus, for example, the edges 4, 5, 2, 1 represent the same circuit as above.



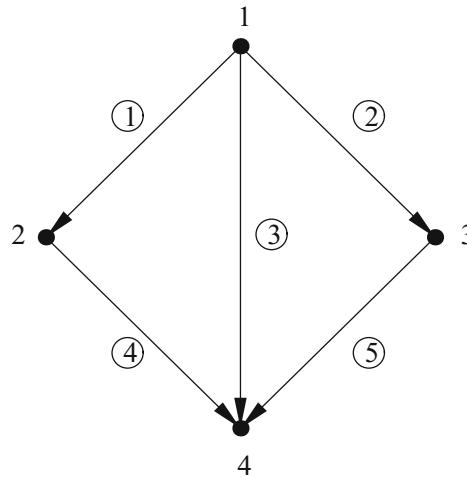
**Figure 2.8.** A Simple Graph.



**Figure 2.9.** Digraphs.

In electrical circuits, one is interested in measuring currents and voltage drops along the wires in the network represented by the graph. Both of these quantities have a direction, and therefore we need to specify an orientation on each edge in order to quantify how the current moves along the wire. The orientation will be fixed by specifying the vertex the edge “starts” at, and the vertex it “ends” at. Once we assign a direction to an edge, a current along that wire will be positive if it moves in the same direction, i.e., goes from the starting vertex to the ending one, and negative if it moves in the opposite direction. The direction of the edge does *not* dictate the direction of the current — it just fixes what directions positive and negative values of current represent. A graph with directed edges is known as a *directed graph*, or *digraph* for short. The edge directions are represented by arrows; examples of digraphs can be seen in [Figure 2.9](#). Again, the underlying graph is always assumed to be simple. For example, at any instant in time, the internet can be viewed as a gigantic digraph, in which each vertex represents a web page, and each edge represents an existing link from one page to another.

Consider a digraph  $D$  consisting of  $n$  vertices connected by  $m$  edges. The *incidence matrix* associated with  $D$  is an  $m \times n$  matrix  $A$  whose rows are indexed by the edges and whose columns are indexed by the vertices. If edge  $k$  starts at vertex  $i$  and ends at vertex  $j$ , then row  $k$  of the incidence matrix will have  $+1$  in its  $(k, i)$  entry and  $-1$  in its  $(k, j)$  entry; all other entries in the row are zero. Thus, our convention is that  $+1$  represents the



**Figure 2.10.** A Simple Digraph.

outgoing vertex at which the edge starts and  $-1$  the incoming vertex at which it ends.

A simple example is the digraph in [Figure 2.10](#), which consists of five edges joined at four different vertices. Its  $5 \times 4$  incidence matrix is

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \quad (2.46)$$

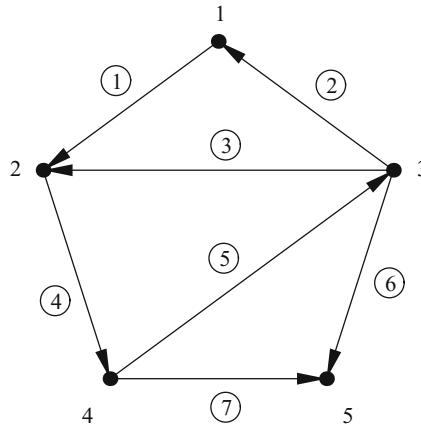
Thus the first row of  $A$  tells us that the first edge starts at vertex 1 and ends at vertex 2. Similarly, row 2 says that the second edge goes from vertex 1 to vertex 3, and so on. Clearly, one can completely reconstruct any digraph from its incidence matrix.

**Example 2.50.** The matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad (2.47)$$

qualifies as an incidence matrix of a simple graph because each row contains a single  $+1$ , a single  $-1$ , and the other entries are 0; moreover, to ensure simplicity, no two rows are identical or  $-1$  times each other. Let us construct the digraph corresponding to  $A$ . Since  $A$  has five columns, there are five vertices in the digraph, which we label by the numbers 1, 2, 3, 4, 5. Since it has seven rows, there are 7 edges. The first row has its  $+1$  in column 1 and its  $-1$  in column 2, and so the first edge goes from vertex 1 to vertex 2. Similarly, the second edge corresponds to the second row of  $A$  and so goes from vertex 3 to vertex 1. The third row of  $A$  indicates an edge from vertex 3 to vertex 2; and so on. In this manner, we construct the digraph drawn in [Figure 2.11](#).

The incidence matrix serves to encode important geometric information about the digraph it represents. In particular, its kernel and cokernel have topological significance.



**Figure 2.11.** Another Digraph.

For example, the kernel of the incidence matrix (2.47) is spanned by the single vector  $\mathbf{z} = (1, 1, 1, 1, 1)^T$ , and represents the fact that the sum of the entries in any given row of  $A$  is zero. This observation holds in general for connected digraphs.

**Proposition 2.51.** If  $A$  is the incidence matrix for a connected digraph, then  $\ker A$  is one-dimensional, with basis  $\mathbf{z} = (1, 1, \dots, 1)^T$ .

*Proof:* If edge  $k$  connects vertex  $i$  to vertex  $j$ , then the  $k^{\text{th}}$  equation in  $A\mathbf{z} = \mathbf{0}$  is  $z_i - z_j = 0$ , or, equivalently,  $z_i = z_j$ . The same equality holds, by a simple induction, if the vertices  $i$  and  $j$  are connected by a path. Therefore, if  $D$  is connected, then all the entries of  $\mathbf{z}$  are equal, and the result follows. *Q.E.D.*

**Remark.** In general,  $\dim \ker A$  equals the number of connected components in the digraph  $D$ . See Exercise 2.6.12.

Applying the Fundamental Theorem 2.49, we immediately deduce the following:

**Corollary 2.52.** If  $A$  is the incidence matrix for a connected digraph with  $n$  vertices, then  $\text{rank } A = n - 1$ .

Next, let us look at the cokernel of an incidence matrix. Consider the particular example (2.46) corresponding to the digraph in [Figure 2.10](#). We need to compute the kernel of the transposed incidence matrix

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & -1 \end{pmatrix}. \quad (2.48)$$

Solving the homogeneous system  $A^T \mathbf{y} = \mathbf{0}$  by Gaussian Elimination, we discover that  $\text{coker } A = \ker A^T$  is spanned by the two vectors

$$\mathbf{y}_1 = (1, 0, -1, 1, 0)^T, \quad \mathbf{y}_2 = (0, 1, -1, 0, 1)^T.$$

Each of these vectors represents a *circuit* in the digraph. Keep in mind that their entries are indexed by the edges, so a nonzero entry indicates the direction to traverse the corresponding edge. For example,  $\mathbf{y}_1$  corresponds to the circuit that starts out along edge 1,

then goes along edge 4 and finishes by going along edge 3 in the reverse direction, which is indicated by the minus sign in its third entry. Similarly,  $\mathbf{y}_2$  represents the circuit consisting of edge 2, followed by edge 5, and then edge 3, backwards. The fact that  $\mathbf{y}_1$  and  $\mathbf{y}_2$  are linearly independent vectors says that the two circuits are “independent”.

The general element of  $\text{coker } A$  is a linear combination  $c_1 \mathbf{y}_1 + c_2 \mathbf{y}_2$ . Certain values of the constants lead to other types of circuits; for example,  $-\mathbf{y}_1$  represents the same circuit as  $\mathbf{y}_1$ , but traversed in the opposite direction. Another example is

$$\mathbf{y}_1 - \mathbf{y}_2 = (1, -1, 0, 1, -1)^T,$$

which represents the square circuit going around the outside of the digraph along edges 1, 4, 5, 2, the fifth and second edges taken in the reverse direction. We can view this circuit as a combination of the two triangular circuits; when we add them together, the middle edge 3 is traversed once in each direction, which effectively “cancels” its contribution. (A similar cancellation occurs in the calculus of line integrals, [2, 78].) Other combinations represent “virtual” circuits; for instance, one can “interpret”  $2\mathbf{y}_1 - \frac{1}{2}\mathbf{y}_2$  as two times around the first triangular circuit plus one-half of the other triangular circuit, taken in the reverse direction — whatever that might mean.

Let us summarize the preceding discussion.

**Theorem 2.53.** Each circuit in a digraph  $D$  is represented by a vector in the cokernel of its incidence matrix  $A$ , whose entries are  $+1$  if the edge is traversed in the correct direction,  $-1$  if in the opposite direction, and  $0$  if the edge is not in the circuit. The dimension of the cokernel of  $A$  equals the number of independent circuits in  $D$ .

**Remark.** A full proof that the cokernel of the incidence matrix of a general digraph has a basis consisting entirely of independent circuits requires a more in depth analysis of the properties of graphs than we can provide in this abbreviated treatment. Full details can be found in [6; §II.3].

The preceding two theorems have an important and remarkable consequence. Suppose  $D$  is a connected digraph with  $m$  edges and  $n$  vertices and  $A$  its  $m \times n$  incidence matrix. Corollary 2.52 implies that  $A$  has rank  $r = n - 1 = n - \dim \ker A$ . On the other hand, Theorem 2.53 tells us that  $l = \dim \text{coker } A$  equals the number of independent circuits in  $D$ . The Fundamental Theorem 2.49 says that  $r = m - l$ . Equating these two formulas for the rank, we obtain  $r = n - 1 = m - l$ , or  $n + l = m + 1$ . This celebrated result is known as *Euler’s formula* for graphs, first discovered by the extraordinarily prolific and influential eighteenth-century Swiss mathematician Leonhard Euler<sup>†</sup>.

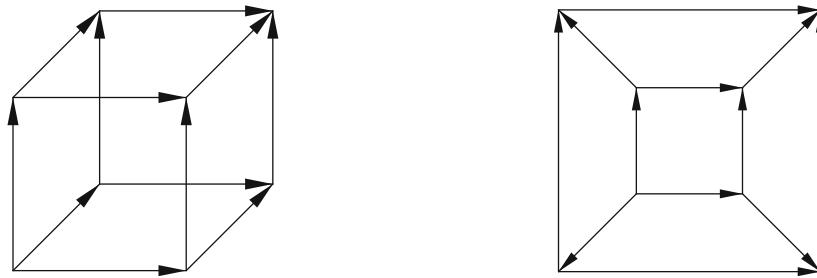
**Theorem 2.54.** If  $G$  is a connected graph, then

$$\# \text{ vertices} + \# \text{ independent circuits} = \# \text{ edges} + 1. \quad (2.49)$$

**Remark.** If the graph is *planar*, meaning that it can be graphed in the plane without any edges crossing over each other, then the number of independent circuits is equal to the number of “holes”, i.e., the number of distinct polygonal regions bounded by the edges of the graph. For example, the pentagonal digraph in Figure 2.11 bounds three triangles, and so has three independent circuits.

---

<sup>†</sup> Pronounced “Oiler”. Euler spent most of his career in Russia and Germany.



**Figure 2.12.** A Cubical Digraph.

**Example 2.55.** Consider the graph corresponding to the edges of a cube, as illustrated in Figure 2.12, where the second figure represents the same graph squashed down onto a plane. The graph has 8 vertices and 12 edges. Euler's formula (3.92) tells us that there are 5 independent circuits. These correspond to the interior square and four trapezoids in the planar version of the digraph, and hence to circuits around 5 of the 6 faces of the cube. The “missing” face does indeed define a circuit, but it can be represented as the sum of the other five circuits, and so is not independent. In Exercise 2.6.6, the reader is asked to write out the incidence matrix for the cubical digraph and explicitly identify the basis of its kernel with the circuits.

Further development of the many remarkable connections between graph theory and linear algebra will be developed in the later chapters. The applications to very large graphs, e.g., with millions or billions of vertices, is playing an increasingly important role in modern computer science and data analysis. One example is the dominant internet search engine run by Google, which is based on viewing the entire internet as a gigantic (time-dependent) digraph. The vertices are the web pages, and a directed edge represents a link from one web page to another. (The resulting digraph is not simple according to our definition, since web pages can link in both directions.) Ranking web pages by importance during a search relies on analyzing the internet digraph; see Section 9.3 for further details.

## Exercises

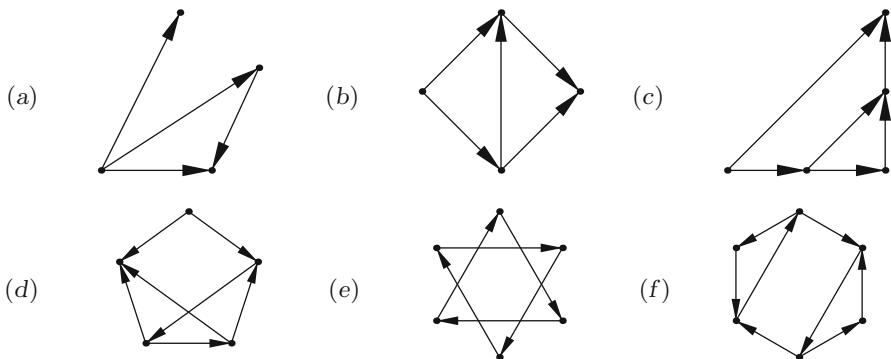
2.6.1. (a) Draw the graph corresponding to the  $6 \times 7$  incidence matrix whose nonzero  $(i, j)$  entries equal 1 if  $j = i$  and  $-1$  if  $j = i + 1$ , for  $i = 1$  to 6. (b) Find a basis for its kernel and cokernel. (c) How many circuits are in the digraph?

2.6.2. Draw the digraph represented by the following incidence matrices:

$$(a) \begin{pmatrix} -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 1 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 & 0 \end{pmatrix},$$

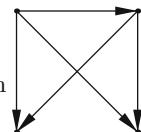
$$(d) \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}, \quad (e) \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

2.6.3. Write out the incidence matrix of the following digraphs.

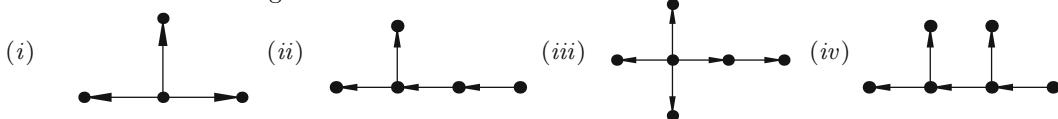


2.6.4. For each of the digraphs in Exercise 2.6.3, see whether you can predict a collection of independent circuits. Verify your prediction by constructing a suitable basis of the cokernel of the incidence matrix and identifying each basis vector with a circuit.

- ◇ 2.6.5. (a) Write down the incidence matrix  $A$  for the indicated digraph.  
 (b) What is the rank of  $A$ ? (c) Determine the dimensions of its four fundamental subspaces. (d) Find a basis for its kernel and cokernel.  
 (e) Determine explicit conditions on vectors  $\mathbf{b}$  that guarantee that the system  $A\mathbf{x} = \mathbf{b}$  has a solution. (f) Write down a specific nonzero vector  $\mathbf{b}$  that satisfies your conditions, and then find all possible solutions.
- ◇ 2.6.6. (a) Write out the incidence matrix for the cubical digraph and identify the basis of its cokernel with the circuits. (b) Find three circuits that do not correspond to any of your basis elements, and express them as a linear combination of the basis circuit vectors.
- ◇ 2.6.7. Write out the incidence matrix for the other Platonic solids: (a) tetrahedron, (b) octahedron, (c) dodecahedron, and (d) icosahedron. (You will need to choose an orientation for the edges.) Show that, in each case, the number of independent circuits equals the number of faces minus 1.
- ◇ 2.6.8. Prove that a graph with  $n$  vertices and  $n$  edges must have at least one circuit.



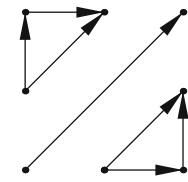
- ◇ 2.6.9. A connected graph is called a *tree* if it has no circuits. (a) Find the incidence matrix for each of the following directed trees:



- (b) Draw all distinct trees with 4 vertices. Assign a direction to the edges, and write down the corresponding incidence matrices. (c) Prove that a connected graph on  $n$  vertices is a tree if and only if it has precisely  $n - 1$  edges.

- ◇ 2.6.10. A *complete graph*  $G_n$  on  $n$  vertices has one edge joining every distinct pair of vertices.  
 (a) Draw  $G_3$ ,  $G_4$  and  $G_5$ . (b) Choose an orientation for each edge and write out the resulting incidence matrix of each digraph. (c) How many edges does  $G_n$  have? (d) How many independent circuits?
- ◇ 2.6.11. The *complete bipartite digraph*  $G_{m,n}$  is based on two disjoint sets of, respectively,  $m$  and  $n$  vertices. Each vertex in the first set is connected to each vertex in the second set by a single edge. (a) Draw  $G_{2,3}$ ,  $G_{2,4}$ , and  $G_{3,3}$ . (b) Write the incidence matrix of each digraph. (c) How many edges does  $G_{m,n}$  have? (d) How many independent circuits?

- ◇ 2.6.12. (a) Construct the incidence matrix  $A$  for the disconnected digraph  $D$  in the figure. (b) Verify that  $\dim \ker A = 3$ , which is the same as the number of connected components, meaning the maximal connected subgraphs in  $D$ . (c) Can you assign an interpretation to your basis for  $\ker A$ ? (d) Try proving the general statement that  $\dim \ker A$  equals the number of connected components in the digraph  $D$ .



- 2.6.13. How does altering the direction of the edges of a digraph affect its incidence matrix? The cokernel of its incidence matrix? Can you realize this operation by matrix multiplication?

- ◇ 2.6.14. (a) Explain why two digraphs are equivalent under relabeling of vertices and edges if and only if their incidence matrices satisfy  $PAQ = B$ , where  $P, Q$  are permutation matrices. (b) Decide which of the following incidence matrices produce the equivalent digraphs:

$$(i) \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad (ii) \begin{pmatrix} 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 1 \end{pmatrix}, \quad (iii) \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

$$(iv) \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \quad (v) \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}, \quad (vi) \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{pmatrix}.$$

- (c) How are the cokernels of equivalent incidence matrices related?

- 2.6.15. *True or false:* If  $A$  and  $B$  are incidence matrices of the same size and  $\text{coker } A = \text{coker } B$ , then the corresponding digraphs are equivalent.

- ◇ 2.6.16. (a) Explain why the incidence matrix for a disconnected graph can be written in block diagonal matrix form  $A = \begin{pmatrix} B & O \\ O & C \end{pmatrix}$  under an appropriate labeling of the vertices.  
 (b) Show how to label the vertices of the digraph in Exercise 2.6.3e so that its incidence matrix is in block form.
-



# Chapter 3

# Inner Products and Norms

The geometry of Euclidean space is founded on the familiar properties of length and angle. The abstract concept of a norm on a vector space formalizes the geometrical notion of the length of a vector. In Euclidean geometry, the angle between two vectors is specified by their dot product, which is itself formalized by the abstract concept of an inner product. Inner products and norms lie at the heart of linear (and nonlinear) analysis, in both finite-dimensional vector spaces and infinite-dimensional function spaces. A vector space equipped with an inner product and its associated norm is known as an inner product space. It is impossible to overemphasize their importance for theoretical developments, practical applications, and the design of numerical solution algorithms.

Mathematical analysis relies on the exploitation of inequalities. The most fundamental is the Cauchy–Schwarz inequality, which is valid in every inner product space. The more familiar triangle inequality for the associated norm is then derived as a simple consequence. Not every norm comes from an inner product, and, in such cases, the triangle inequality becomes part of the general definition. Both inequalities retain their validity in both finite-dimensional and infinite-dimensional vector spaces. Indeed, their abstract formulation exposes the key ideas behind the proof, avoiding all distracting particularities appearing in the explicit formulas.

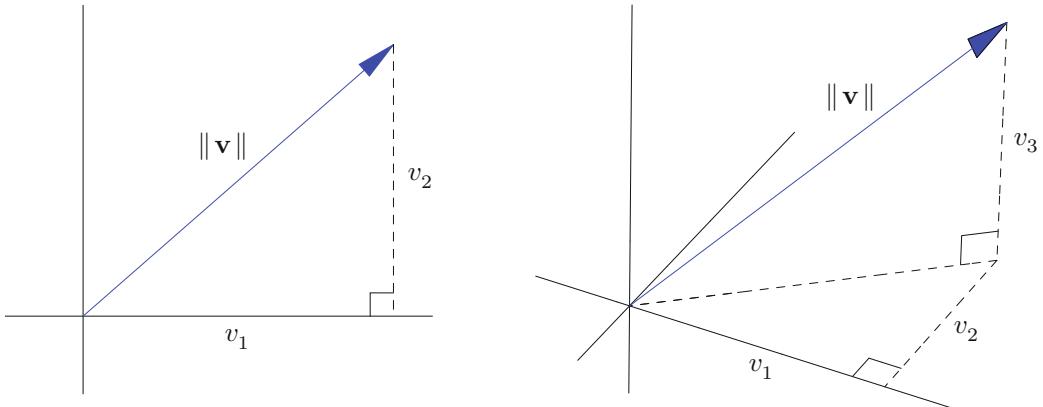
The characterization of general inner products on Euclidean space will lead us to the noteworthy class of positive definite matrices. Positive definite matrices appear in a wide variety of applications, including minimization, least squares, data analysis and statistics, as well as, for example, mechanical systems, electrical circuits, and the differential equations describing both static and dynamical processes. The test for positive definiteness relies on Gaussian Elimination, and we can reinterpret the resulting matrix factorization as the algebraic process of completing the square for the associated quadratic form. In applications, positive definite matrices most often arise as Gram matrices, whose entries are formed by taking inner products between selected elements of an inner product space.

So far, we have focussed our attention on real vector spaces. Complex numbers, vectors, and functions also arise in numerous applications, and so, in the final section, we take the opportunity to formally introduce complex vector spaces. Most of the theory proceeds in direct analogy with the real version, but the notions of inner product and norm on complex vector spaces require some thought. Applications of complex vector spaces and their inner products are of particular significance in Fourier analysis, signal processing, and partial differential equations, [61], and they play an absolutely essential role in modern quantum mechanics, [54].

## 3.1 Inner Products

The most basic example of an inner product is the familiar *dot product*

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n = \sum_{i=1}^n v_i w_i, \quad (3.1)$$



**Figure 3.1.** The Euclidean Norm in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ .

between (column) vectors  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$ ,  $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$ , both lying in the Euclidean space  $\mathbb{R}^n$ . A key observation is that the dot product (3.1) is equal to the matrix product

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w} = (v_1 \ v_2 \ \dots \ v_n) \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \quad (3.2)$$

between the row vector  $\mathbf{v}^T$  and the column vector  $\mathbf{w}$ .

The dot product is the cornerstone of Euclidean geometry. The key fact is that the dot product of a vector with itself,

$$\mathbf{v} \cdot \mathbf{v} = v_1^2 + v_2^2 + \dots + v_n^2,$$

is the sum of the squares of its entries, and hence, by the classical Pythagorean Theorem, equals the square of its length; see Figure 3.1. Consequently, the *Euclidean norm* or *length* of a vector is found by taking the square root:

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}. \quad (3.3)$$

Note that every nonzero vector,  $\mathbf{v} \neq \mathbf{0}$ , has positive Euclidean norm,  $\|\mathbf{v}\| > 0$ , while only the zero vector has zero norm:  $\|\mathbf{v}\| = 0$  if and only if  $\mathbf{v} = \mathbf{0}$ . The elementary properties of dot product and Euclidean norm serve to inspire the abstract definition of more general inner products.

**Definition 3.1.** An *inner product* on the real vector space  $V$  is a pairing that takes two vectors  $\mathbf{v}, \mathbf{w} \in V$  and produces a real number  $\langle \mathbf{v}, \mathbf{w} \rangle \in \mathbb{R}$ . The inner product is required to satisfy the following three axioms for all  $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ , and scalars  $c, d \in \mathbb{R}$ .

$$(i) \text{ Bilinearity:} \quad \begin{aligned} \langle c\mathbf{u} + d\mathbf{v}, \mathbf{w} \rangle &= c\langle \mathbf{u}, \mathbf{w} \rangle + d\langle \mathbf{v}, \mathbf{w} \rangle, \\ \langle \mathbf{u}, c\mathbf{v} + d\mathbf{w} \rangle &= c\langle \mathbf{u}, \mathbf{v} \rangle + d\langle \mathbf{u}, \mathbf{w} \rangle. \end{aligned} \quad (3.4)$$

$$(ii) \text{ Symmetry:} \quad \langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle. \quad (3.5)$$

$$(iii) \text{ Positivity:} \quad \langle \mathbf{v}, \mathbf{v} \rangle > 0 \quad \text{whenever} \quad \mathbf{v} \neq \mathbf{0}, \quad \text{while} \quad \langle \mathbf{0}, \mathbf{0} \rangle = 0. \quad (3.6)$$

A vector space equipped with an inner product is called an *inner product space*. As we shall see, a vector space can admit many different inner products. Verification of the inner

product axioms for the Euclidean dot product is straightforward, and left as an exercise for the reader.

Given an inner product, the associated *norm* of a vector  $\mathbf{v} \in V$  is defined as the positive square root of the inner product of the vector with itself:

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}. \quad (3.7)$$

The positivity axiom implies that  $\|\mathbf{v}\| \geq 0$  is real and non-negative, and equals 0 if and only if  $\mathbf{v} = \mathbf{0}$  is the zero vector.

**Example 3.2.** While certainly the most common inner product on  $\mathbb{R}^2$ , the dot product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2$$

is by no means the only possibility. A simple example is provided by the *weighted inner product*

$$\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1 w_1 + 5v_2 w_2, \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \quad (3.8)$$

Let us verify that this formula does indeed define an inner product. The symmetry axiom (3.5) is immediate. Moreover,

$$\begin{aligned} \langle c\mathbf{u} + d\mathbf{v}, \mathbf{w} \rangle &= 2(cu_1 + dv_1)w_1 + 5(cu_2 + dv_2)w_2 \\ &= c(2u_1 w_1 + 5u_2 w_2) + d(2v_1 w_1 + 5v_2 w_2) = c\langle \mathbf{u}, \mathbf{w} \rangle + d\langle \mathbf{v}, \mathbf{w} \rangle, \end{aligned}$$

which verifies the first bilinearity condition; the second follows by a very similar computation. (Or, one can use the symmetry axiom to deduce the second bilinearity identity from the first; see Exercise 3.1.9.) Moreover,  $\langle \mathbf{0}, \mathbf{0} \rangle = 0$ , while

$$\langle \mathbf{v}, \mathbf{v} \rangle = 2v_1^2 + 5v_2^2 > 0 \quad \text{whenever} \quad \mathbf{v} \neq \mathbf{0},$$

since at least one of the summands is strictly positive. This establishes (3.8) as a legitimate inner product on  $\mathbb{R}^2$ . The associated *weighted norm*  $\|\mathbf{v}\| = \sqrt{2v_1^2 + 5v_2^2}$  defines an alternative, “non-Pythagorean” notion of length of vectors and distance between points in the plane.

A less evident example of an inner product on  $\mathbb{R}^2$  is provided by the expression

$$\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 - v_1 w_2 - v_2 w_1 + 4v_2 w_2. \quad (3.9)$$

Bilinearity is verified in the same manner as before, and symmetry is immediate. Positivity is ensured by noticing that the expression

$$\langle \mathbf{v}, \mathbf{v} \rangle = v_1^2 - 2v_1 v_2 + 4v_2^2 = (v_1 - v_2)^2 + 3v_2^2 \geq 0$$

is always non-negative, and, moreover, is equal to zero if and only if  $v_1 - v_2 = 0$ ,  $v_2 = 0$ , i.e., only when  $v_1 = v_2 = 0$  and so  $\mathbf{v} = \mathbf{0}$ . We conclude that (3.9) defines yet another inner product on  $\mathbb{R}^2$ , with associated norm

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{v_1^2 - 2v_1 v_2 + 4v_2^2}.$$

The second example (3.8) is a particular case of a general class of inner products.

**Example 3.3.** Let  $c_1, \dots, c_n > 0$  be a set of *positive* numbers. The corresponding *weighted inner product* and *weighted norm* on  $\mathbb{R}^n$  are defined by

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^n c_i v_i w_i, \quad \|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\sum_{i=1}^n c_i v_i^2}. \quad (3.10)$$

The numbers  $c_i > 0$  are the *weights*. Observe that the larger the weight  $c_i$ , the more the  $i^{\text{th}}$  coordinate of  $\mathbf{v}$  contributes to the norm. Weighted norms are particularly relevant in statistics and data fitting, [43, 87], when one wants to emphasize the importance of certain measurements and de-emphasize others; this is done by assigning appropriate weights to the different components of the data vector  $\mathbf{v}$ . Section 5.4, on least squares approximation methods, will contain further details.

---

## Exercises

3.1.1. Prove that the formula  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 - v_1 w_2 - v_2 w_1 + b v_2 w_2$  defines an inner product on  $\mathbb{R}^2$  if and only if  $b > 1$ .

3.1.2. Which of the following formulas for  $\langle \mathbf{v}, \mathbf{w} \rangle$  define inner products on  $\mathbb{R}^2$ ?

- (a)  $2v_1 w_1 + 3v_2 w_2$ , (b)  $v_1 w_2 + v_2 w_1$ , (c)  $(v_1 + v_2)(w_1 + w_2)$ , (d)  $v_1^2 w_1^2 + v_2^2 w_2^2$ ,
- (e)  $\sqrt{v_1^2 + v_2^2} \sqrt{w_1^2 + w_2^2}$ , (f)  $2v_1 w_1 + (v_1 - v_2)(w_1 - w_2)$ ,
- (g)  $4v_1 w_1 - 2v_1 w_2 - 2v_2 w_1 + 4v_2 w_2$ .

3.1.3. Show that  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + v_1 w_2 + v_2 w_1 + v_2 w_2$  does *not* define an inner product on  $\mathbb{R}^2$ .

3.1.4. Prove that each of the following formulas for  $\langle \mathbf{v}, \mathbf{w} \rangle$  defines an inner product on  $\mathbb{R}^3$ .

Verify all the inner product axioms in careful detail:

- (a)  $v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$ , (b)  $4v_1 w_1 + 2v_1 w_2 + 2v_2 w_1 + 4v_2 w_2 + v_3 w_3$ ,
- (c)  $2v_1 w_1 - 2v_1 w_2 - 2v_2 w_1 + 3v_2 w_2 - v_2 w_3 - v_3 w_2 + 2v_3 w_3$ .

3.1.5. The *unit circle* for an inner product on  $\mathbb{R}^2$  is defined as the set of all vectors of unit length:  $\|\mathbf{v}\| = 1$ . Graph the unit circles for (a) the Euclidean inner product, (b) the weighted inner product (3.8), (c) the non-standard inner product (3.9). (d) Prove that cases (b) and (c) are, in fact, both ellipses.

◇ 3.1.6. (a) Explain why the formula for the Euclidean norm in  $\mathbb{R}^2$  follows from the Pythagorean Theorem. (b) How do you use the Pythagorean Theorem to justify the formula for the Euclidean norm in  $\mathbb{R}^3$ ? *Hint:* Look at Figure 3.1.

◇ 3.1.7. Prove that the norm on an inner product space satisfies  $\|c\mathbf{v}\| = |c| \|\mathbf{v}\|$  for every scalar  $c$  and vector  $\mathbf{v}$ .

3.1.8. Prove that  $\langle a\mathbf{v} + b\mathbf{w}, c\mathbf{v} + d\mathbf{w} \rangle = ac\|\mathbf{v}\|^2 + (ad + bc)\langle \mathbf{v}, \mathbf{w} \rangle + bd\|\mathbf{w}\|^2$ .

◇ 3.1.9. Prove that the second bilinearity formula (3.4) is a consequence of the first and the other two inner product axioms.

◇ 3.1.10. Let  $V$  be an inner product space. (a) Prove that  $\langle \mathbf{x}, \mathbf{v} \rangle = 0$  for all  $\mathbf{v} \in V$  if and only if  $\mathbf{x} = \mathbf{0}$ . (b) Prove that  $\langle \mathbf{x}, \mathbf{v} \rangle = \langle \mathbf{y}, \mathbf{v} \rangle$  for all  $\mathbf{v} \in V$  if and only if  $\mathbf{x} = \mathbf{y}$ .  
(c) Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis for  $V$ . Prove that  $\langle \mathbf{x}, \mathbf{v}_i \rangle = \langle \mathbf{y}, \mathbf{v}_i \rangle$ ,  $i = 1, \dots, n$ , if and only if  $\mathbf{x} = \mathbf{y}$ .

◇ 3.1.11. Prove that  $\mathbf{x} \in \mathbb{R}^n$  solves the linear system  $A\mathbf{x} = \mathbf{b}$  if and only if

$$\mathbf{x}^T A^T \mathbf{v} = \mathbf{b}^T \mathbf{v} \quad \text{for all } \mathbf{v} \in \mathbb{R}^m.$$

The latter is known as the *weak formulation* of the linear system, and its generalizations are of great importance in the study of differential equations and numerical analysis, [61].

◇ 3.1.12. (a) Prove the identity

$$\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{4} \left( \|\mathbf{u} + \mathbf{v}\|^2 - \|\mathbf{u} - \mathbf{v}\|^2 \right), \quad (3.11)$$

which allows one to reconstruct an inner product from its norm. (b) Use (3.11) to find the inner product on  $\mathbb{R}^2$  corresponding to the norm  $\|\mathbf{v}\| = \sqrt{v_1^2 - 3v_1 v_2 + 5v_2^2}$ .

3.1.13. (a) Show that, for all vectors  $\mathbf{x}$  and  $\mathbf{y}$  in an inner product space,

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2).$$

(b) Interpret this result pictorially for vectors in  $\mathbb{R}^2$  under the Euclidean norm.

3.1.14. Suppose  $\mathbf{u}, \mathbf{v}$  satisfy  $\|\mathbf{u}\| = 3$ ,  $\|\mathbf{u} + \mathbf{v}\| = 4$ , and  $\|\mathbf{u} - \mathbf{v}\| = 6$ . What must  $\|\mathbf{v}\|$  equal? Does your answer depend upon which norm is being used?

3.1.15. Let  $A$  be any  $n \times n$  matrix. Prove that the dot product identity  $\mathbf{v} \cdot (A\mathbf{w}) = (A^T\mathbf{v}) \cdot \mathbf{w}$  is valid for all vectors  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ .

◊ 3.1.16. Prove that  $A = A^T$  is a symmetric  $n \times n$  matrix if and only if  $(A\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\mathbf{w})$  for all  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ .

3.1.17. Prove that  $\langle A, B \rangle = \text{tr}(A^T B)$  defines an inner product on the vector space  $\mathcal{M}_{n \times n}$  of real  $n \times n$  matrices.

3.1.18. Suppose  $\langle \mathbf{v}, \mathbf{w} \rangle$  defines an inner product on a vector space  $V$ . Explain why it also defines an inner product on every subspace  $W \subset V$ .

3.1.19. Prove that if  $\langle \mathbf{v}, \mathbf{w} \rangle$  and  $\langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle$  are two inner products on the same vector space  $V$ , then their sum  $\langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle = \langle \mathbf{v}, \mathbf{w} \rangle + \langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle$  defines an inner product on  $V$ .

◊ 3.1.20. Let  $V$  and  $W$  be inner product spaces with respective inner products  $\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle$  and  $\langle\langle \mathbf{w}, \tilde{\mathbf{w}} \rangle\rangle$ . Show that  $\langle\langle (\mathbf{v}, \mathbf{w}), (\tilde{\mathbf{v}}, \tilde{\mathbf{w}}) \rangle\rangle = \langle \mathbf{v}, \tilde{\mathbf{v}} \rangle + \langle\langle \mathbf{w}, \tilde{\mathbf{w}} \rangle\rangle$  for  $\mathbf{v}, \tilde{\mathbf{v}} \in V$ ,  $\mathbf{w}, \tilde{\mathbf{w}} \in W$ , defines an inner product on their Cartesian product  $V \times W$ .

## Inner Products on Function Spaces

Inner products and norms on function spaces lie at the foundation of modern analysis and its applications, particularly Fourier analysis, boundary value problems, ordinary and partial differential equations, and numerical analysis. Let us introduce the most important examples.

**Example 3.4.** Let  $[a, b] \subset \mathbb{R}$  be a bounded closed interval. Consider the vector space  $C^0[a, b]$  consisting of all continuous scalar functions  $f$  defined on the interval  $[a, b]$ . The integral of the product of two continuous functions,

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx, \quad (3.12)$$

defines an inner product on the vector space  $C^0[a, b]$ , as we shall prove below. The associated norm is, according to the basic definition (3.7),

$$\|f\| = \sqrt{\int_a^b f(x)^2 dx}, \quad (3.13)$$

and is known as the  $L^2$  norm of the function  $f$  over the interval  $[a, b]$ . The  $L^2$  inner product and norm of functions can be viewed as the infinite-dimensional function space versions of the dot product and Euclidean norm of vectors in  $\mathbb{R}^n$ . The reason for the name  $L^2$  will become clearer later on.

For example, if we take  $[a, b] = [0, \frac{1}{2}\pi]$ , then the  $L^2$  inner product between  $f(x) = \sin x$  and  $g(x) = \cos x$  is equal to

$$\langle \sin x, \cos x \rangle = \int_0^{\pi/2} \sin x \cos x dx = \frac{1}{2} \sin^2 x \Big|_{x=0}^{\pi/2} = \frac{1}{2}.$$

Similarly, the norm of the function  $\sin x$  is

$$\|\sin x\| = \sqrt{\int_0^{\pi/2} (\sin x)^2 dx} = \sqrt{\frac{\pi}{4}}.$$

One must always be careful when evaluating function norms. For example, the constant function  $c(x) \equiv 1$  has norm

$$\|1\| = \sqrt{\int_0^{\pi/2} 1^2 dx} = \sqrt{\frac{\pi}{2}},$$

not 1 as you might have expected. We also note that the value of the norm depends upon which interval the integral is taken over. For instance, on the longer interval  $[0, \pi]$ ,

$$\|1\| = \sqrt{\int_0^{\pi} 1^2 dx} = \sqrt{\pi}.$$

Thus, when dealing with the  $L^2$  inner product or norm, one must always be careful to specify the function space, or, equivalently, the interval on which it is being evaluated.

Let us prove that formula (3.12) does, indeed, define an inner product. First, we need to check that  $\langle f, g \rangle$  is well defined. This follows because the product  $f(x)g(x)$  of two continuous functions is also continuous, and hence its integral over a bounded interval is defined and finite. The symmetry requirement is immediate:

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx = \int_a^b g(x) f(x) dx = \langle g, f \rangle,$$

because multiplication of functions is commutative. The first bilinearity axiom

$$\langle c f + d g, h \rangle = c \langle f, h \rangle + d \langle g, h \rangle$$

amounts to the following elementary integral identity

$$\int_a^b [c f(x) + d g(x)] h(x) dx = c \int_a^b f(x) h(x) dx + d \int_a^b g(x) h(x) dx,$$

valid for arbitrary continuous functions  $f, g, h$  and scalars (constants)  $c, d$ . The second bilinearity axiom is proved similarly; alternatively, one can use symmetry to deduce it from the first as in Exercise 3.1.9. Finally, positivity requires that

$$\|f\|^2 = \langle f, f \rangle = \int_a^b f(x)^2 dx \geq 0.$$

This is clear because  $f(x)^2 \geq 0$ , and the integral of a nonnegative function is nonnegative. Moreover, since the function  $f(x)^2$  is continuous and nonnegative, its integral will vanish,  $\int_a^b f(x)^2 dx = 0$ , if and only if  $f(x) \equiv 0$  is the zero function, cf. Exercise 3.1.29. This completes the proof that (3.12) defines a bona fide inner product on the space  $C^0[a, b]$ .

**Remark.** The  $L^2$  inner product formula can also be applied to more general functions, but we have restricted our attention to continuous functions in order to avoid certain technical complications. The most general function space admitting this inner product is known

as *Hilbert space*, which forms the basis of much of modern analysis, function theory, and Fourier analysis, as well as providing the theoretical setting for all of quantum mechanics, [54]. Unfortunately, we cannot provide the mathematical details of the Hilbert space construction, since it requires that you be familiar with measure theory and the Lebesgue integral. See [61] for a basic introduction and [19, 68, 77] for the fully rigorous theory.

**Warning.** One needs to be extremely careful when trying to extend the  $L^2$  inner product to other spaces of functions. Indeed, there are *nonzero* discontinuous functions with *zero* “ $L^2$  norm”. For example, the function

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & \text{otherwise,} \end{cases} \quad \text{satisfies} \quad \|f\|^2 = \int_{-1}^1 f(x)^2 dx = 0, \quad (3.14)$$

because every function that is zero except at finitely many (or even countably many) points has zero integral.

The  $L^2$  inner product is but one of a vast number of possible inner products on function spaces. For example, one can also define weighted inner products on the space  $C^0[a, b]$ . The weighting along the interval is specified by a (continuous) positive scalar function  $w(x) > 0$ . The corresponding *weighted inner product* and *norm* are

$$\langle f, g \rangle = \int_a^b f(x) g(x) w(x) dx, \quad \|f\| = \sqrt{\int_a^b f(x)^2 w(x) dx}. \quad (3.15)$$

The verification of the inner product axioms in this case is left as an exercise for the reader. As in the finite-dimensional version, weighted inner products are often used in statistics and data analysis, [20, 43, 87].

## Exercises

3.1.21. For each of the given pairs of functions in  $C^0[0, 1]$ , find their  $L^2$  inner product

$$\langle f, g \rangle \text{ and their } L^2 \text{ norms } \|f\|, \|g\|: (a) f(x) = 1, g(x) = x; (b) f(x) = \cos 2\pi x, g(x) = \sin 2\pi x; (c) f(x) = x, g(x) = e^x; (d) f(x) = (x+1)^2, g(x) = \frac{1}{x+1}.$$

3.1.22. Let  $f(x) = x, g(x) = 1 + x^2$ . Compute  $\langle f, g \rangle$ ,  $\|f\|$ , and  $\|g\|$  for (a) the  $L^2$  inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) dx$ ; (b) the  $L^2$  inner product  $\langle f, g \rangle = \int_{-1}^1 f(x) g(x) dx$ ; (c) the weighted inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) x dx$ .

3.1.23. Which of the following formulas for  $\langle f, g \rangle$  define inner products on the space

$$C^0[-1, 1]? \quad (a) \int_{-1}^1 f(x) g(x) e^{-x} dx, \quad (b) \int_{-1}^1 f(x) g(x) x dx, \\ (c) \int_{-1}^1 f(x) g(x) (x+2) dx, \quad (d) \int_{-1}^1 f(x) g(x) x^2 dx.$$

3.1.24. Prove that  $\langle f, g \rangle = \int_0^1 f(x) g(x) dx$  does *not* define an inner product on the vector space  $C^0[-1, 1]$ . Explain why this does not contradict the fact that it defines an inner product on the vector space  $C^0[0, 1]$ . Does it define an inner product on the subspace  $\mathcal{P}^{(n)} \subset C^0[-1, 1]$  consisting of all polynomial functions?

3.1.25. Does either of the following define an inner product on  $C^0[0, 1]$ ?

(a)  $\langle f, g \rangle = f(0)g(0) + f(1)g(1)$ , (b)  $\langle f, g \rangle = f(0)g(0) + f(1)g(1) + \int_0^1 f(x)g(x) dx$ .

3.1.26. Let  $f(x)$  be a function, and  $\|f\|$  its  $L^2$  norm on  $[a, b]$ . Is  $\|f^2\| = \|f\|^2$ ? If yes, prove the statement. If no, give a counterexample.

◇ 3.1.27. Prove that  $\langle f, g \rangle = \int_a^b [f(x)g(x) + f'(x)g'(x)] dx$  defines an inner product on the space  $C^1[a, b]$  of continuously differentiable functions on the interval  $[a, b]$ . Write out the corresponding norm, known as the *Sobolev  $H^1$  norm*; it and its generalizations play an extremely important role in advanced mathematical analysis, [49].

3.1.28. Let  $V = C^1[-1, 1]$  denote the vector space of continuously differentiable functions for  $-1 \leq x \leq 1$ . (a) Does the expression  $\langle f, g \rangle = \int_{-1}^1 f'(x)g'(x) dx$  define an inner product on  $V$ ? (b) Answer the same question for the subspace  $W = \{f \in V \mid f(0) = 0\}$  consisting of all continuously differentiable functions that vanish at 0.

◇ 3.1.29. (a) Let  $h(x) \geq 0$  be a continuous, non-negative function defined on an interval  $[a, b]$ .

Prove that  $\int_a^b h(x) dx = 0$  if and only if  $h(x) \equiv 0$ . *Hint:* Use the fact that  $\int_c^d h(x) dx > 0$  if  $h(x) > 0$  for  $c \leq x \leq d$ . (b) Give an example that shows that this result is not valid if  $h$  is allowed to be discontinuous.

◇ 3.1.30. (a) Prove the inner product axioms for the weighted inner product (3.15), assuming  $w(x) > 0$  for all  $a \leq x \leq b$ . (b) Explain why it does *not* define an inner product if  $w$  is continuous and  $w(x_0) < 0$  for some  $x_0 \in [a, b]$ . (c) If  $w(x) \geq 0$  for  $a \leq x \leq b$ , does (3.15) define an inner product? *Hint:* Your answer may depend upon  $w(x)$ .

◇ 3.1.31. Let  $\Omega \subset \mathbb{R}^2$  be a closed bounded subset. Let  $C^0(\Omega)$  denote the vector space consisting of all continuous, bounded real-valued functions  $f(x, y)$  defined for  $(x, y) \in \Omega$ . (a) Prove that if  $f(x, y) \geq 0$  is continuous and  $\iint_{\Omega} f(x, y) dx dy = 0$ , then  $f(x, y) \equiv 0$ . *Hint:* Mimic Exercise 3.1.29. (b) Use this result to prove that

$$\langle f, g \rangle = \iint_{\Omega} f(x, y)g(x, y) dx dy \quad (3.16)$$

defines an inner product on  $C^0(\Omega)$ , called the  $L^2$  inner product on the domain  $\Omega$ . What is the corresponding norm?

3.1.32. Compute the  $L^2$  inner product (3.16) and norms of the functions  $f(x, y) \equiv 1$  and  $g(x, y) = x^2 + y^2$ , when (a)  $\Omega = \{0 \leq x \leq 1, 0 \leq y \leq 1\}$  is the unit square;

(b)  $\Omega = \{x^2 + y^2 \leq 1\}$  is the unit disk. *Hint:* Use polar coordinates.

◇ 3.1.33. Let  $V$  be the vector space consisting of all continuous, vector-valued functions

$\mathbf{f}(x) = (f_1(x), f_2(x))^T$  defined on the interval  $0 \leq x \leq 1$ .

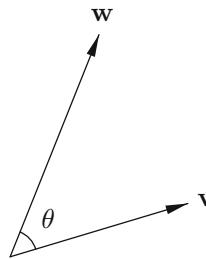
(a) Prove that  $\langle\langle \mathbf{f}, \mathbf{g} \rangle\rangle = \int_0^1 [f_1(x)g_1(x) + f_2(x)g_2(x)] dx$  defines an inner product on  $V$ .

(b) Prove, more generally, that if  $\langle \mathbf{v}, \mathbf{w} \rangle$  is any inner product on  $\mathbb{R}^2$ , then

$\langle\langle \mathbf{f}, \mathbf{g} \rangle\rangle = \int_a^b \langle \mathbf{f}(x), \mathbf{g}(x) \rangle dx$  defines an inner product on  $V$ . (Part (a) corresponds to the dot product.) (c) Use part (b) to prove that

$$\langle\langle \mathbf{f}, \mathbf{g} \rangle\rangle = \int_a^b [f_1(x)g_1(x) - f_1(x)g_2(x) - f_2(x)g_1(x) + 3f_2(x)g_2(x)] dx$$

defines an inner product on  $V$ .



**Figure 3.2.** Angle Between Two Vectors.

## 3.2 Inequalities

There are two absolutely basic inequalities that are valid for *any* inner product space. The first is inspired by the geometric interpretation of the dot product on Euclidean space in terms of the angle between vectors. It is named after two of the founders of modern analysis, the nineteenth-century mathematicians Augustin Cauchy, of France, and Herman Schwarz, of Germany, who established it in the case of the  $L^2$  inner product on function space.<sup>†</sup> The more familiar triangle inequality, that the length of any side of a triangle is bounded by the sum of the lengths of the other two sides, is, in fact, an immediate consequence of the Cauchy–Schwarz inequality, and hence also valid for any norm based on an inner product.

We will present these two inequalities in their most general, abstract form, since this brings their essence into the limelight. Specializing to different inner products and norms on both finite-dimensional and infinite-dimensional vector spaces leads to a wide variety of striking and useful inequalities.

### The Cauchy–Schwarz Inequality

In Euclidean geometry, the dot product between two vectors  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$  can be geometrically characterized by the equation

$$\mathbf{v} \cdot \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \cos \theta, \quad (3.17)$$

where  $\theta = \measuredangle(\mathbf{v}, \mathbf{w})$  measures the angle between the two vectors, as illustrated in [Figure 3.2](#). Since  $|\cos \theta| \leq 1$ , the absolute value of the dot product is bounded by the product of the lengths of the vectors:

$$|\mathbf{v} \cdot \mathbf{w}| \leq \|\mathbf{v}\| \|\mathbf{w}\|.$$

This is the simplest form of the general *Cauchy–Schwarz inequality*. We present a direct algebraic proof that does not rely on the geometrical notions of length and angle and thus demonstrates its universal validity for *any* inner product.

**Theorem 3.5.** Every inner product satisfies the Cauchy–Schwarz inequality

$$|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\| \|\mathbf{w}\|, \quad \text{for all } \mathbf{v}, \mathbf{w} \in V. \quad (3.18)$$

Here,  $\|\mathbf{v}\|$  is the associated norm, while  $|\cdot|$  denotes the absolute value of a real number. Equality holds in (3.18) if and only if  $\mathbf{v}$  and  $\mathbf{w}$  are parallel vectors.

---

<sup>†</sup> Russians also give credit for its discovery to their compatriot Viktor Bunyakovsky, and, indeed, some authors append his name to the inequality.

*Proof:* The case when  $\mathbf{w} = \mathbf{0}$  is trivial, since both sides of (3.18) are equal to 0. Thus, we concentrate on the case when  $\mathbf{w} \neq \mathbf{0}$ . Let  $t \in \mathbb{R}$  be an arbitrary scalar. Using the three inner product axioms, we have

$$\begin{aligned} 0 \leq \|\mathbf{v} + t\mathbf{w}\|^2 &= \langle \mathbf{v} + t\mathbf{w}, \mathbf{v} + t\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{v} \rangle + 2t \langle \mathbf{v}, \mathbf{w} \rangle + t^2 \langle \mathbf{w}, \mathbf{w} \rangle \\ &= \|\mathbf{v}\|^2 + 2t \langle \mathbf{v}, \mathbf{w} \rangle + t^2 \|\mathbf{w}\|^2, \end{aligned} \quad (3.19)$$

with equality holding if and only if  $\mathbf{v} = -t\mathbf{w}$  — which requires  $\mathbf{v}$  and  $\mathbf{w}$  to be parallel vectors. We fix  $\mathbf{v}$  and  $\mathbf{w}$ , and consider the right-hand side of (3.19) as a quadratic function of the scalar variable  $t$ :

$$0 \leq p(t) = at^2 + 2bt + c, \quad \text{where} \quad a = \|\mathbf{w}\|^2, \quad b = \langle \mathbf{v}, \mathbf{w} \rangle, \quad c = \|\mathbf{v}\|^2.$$

To get the maximum mileage out of the fact that  $p(t) \geq 0$ , let us look at where it assumes its minimum, which occurs when its derivative is zero:

$$p'(t) = 2at + 2b = 0, \quad \text{and so} \quad t = -\frac{b}{a} = -\frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{w}\|^2}.$$

Substituting this particular value of  $t$  into (3.19), we obtain

$$0 \leq \|\mathbf{v}\|^2 - 2 \frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{w}\|^2} + \frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{w}\|^2} = \|\mathbf{v}\|^2 - \frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{w}\|^2}.$$

Rearranging this last inequality, we conclude that

$$\frac{\langle \mathbf{v}, \mathbf{w} \rangle^2}{\|\mathbf{w}\|^2} \leq \|\mathbf{v}\|^2, \quad \text{or} \quad \langle \mathbf{v}, \mathbf{w} \rangle^2 \leq \|\mathbf{v}\|^2 \|\mathbf{w}\|^2.$$

Also, as noted above, equality holds if and only if  $\mathbf{v}$  and  $\mathbf{w}$  are parallel. Equality also holds when  $\mathbf{w} = \mathbf{0}$ , which is of course parallel to every vector  $\mathbf{v}$ . Taking the (positive) square root of both sides of the final inequality completes the proof of (3.18). *Q.E.D.*

Given any inner product, we can use the quotient

$$\cos \theta = \frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\| \|\mathbf{w}\|} \quad (3.20)$$

to define the “angle”  $\theta = \measuredangle(\mathbf{v}, \mathbf{w})$  between the vector space elements  $\mathbf{v}, \mathbf{w} \in V$ . The Cauchy–Schwarz inequality tells us that the ratio lies between  $-1$  and  $+1$ , and hence the angle  $\theta$  is well defined modulo  $2\pi$ , and, in fact, unique if we restrict it to lie in the range  $0 \leq \theta \leq \pi$ .

For example, the vectors  $\mathbf{v} = (1, 0, 1)^T$ ,  $\mathbf{w} = (0, 1, 1)^T$  have dot product  $\mathbf{v} \cdot \mathbf{w} = 1$  and norms  $\|\mathbf{v}\| = \|\mathbf{w}\| = \sqrt{2}$ . Hence the Euclidean angle between them is given by

$$\cos \theta = \frac{1}{\sqrt{2} \cdot \sqrt{2}} = \frac{1}{2}, \quad \text{and so} \quad \theta = \measuredangle(\mathbf{v}, \mathbf{w}) = \frac{1}{3}\pi = 1.0471\dots$$

On the other hand, if we adopt the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$ , then  $\mathbf{v} \cdot \mathbf{w} = 3$ ,  $\|\mathbf{v}\| = 2$ ,  $\|\mathbf{w}\| = \sqrt{5}$ , and hence their “weighted” angle becomes

$$\cos \theta = \frac{3}{2\sqrt{5}} = .67082\dots, \quad \text{with} \quad \theta = \measuredangle(\mathbf{v}, \mathbf{w}) = .83548\dots.$$

Thus, the measurement of angle (and length) depends on the choice of an underlying inner product.

Similarly, under the  $L^2$  inner product on the interval  $[0, 1]$ , the “angle”  $\theta$  between the polynomials  $p(x) = x$  and  $q(x) = x^2$  is given by

$$\cos \theta = \frac{\langle x, x^2 \rangle}{\|x\| \|x^2\|} = \frac{\int_0^1 x^3 dx}{\sqrt{\int_0^1 x^2 dx} \sqrt{\int_0^1 x^4 dx}} = \frac{\frac{1}{4}}{\sqrt{\frac{1}{3}} \sqrt{\frac{1}{5}}} = \sqrt{\frac{15}{16}},$$

so that  $\theta = \measuredangle(p, q) = .25268 \dots$  radians.

**Warning.** You should not try to give this notion of angle between functions more significance than the formal definition warrants — it does not correspond to any “angular” properties of their graphs. Also, the value depends on the choice of inner product and the interval upon which it is being computed. For example, if we change to the  $L^2$  inner product on the interval  $[-1, 1]$ , then  $\langle x, x^2 \rangle = \int_{-1}^1 x^3 dx = 0$ . Hence, (3.20) becomes  $\cos \theta = 0$ , so the “angle” between  $x$  and  $x^2$  is now  $\theta = \measuredangle(p, q) = \frac{1}{2}\pi$ .

## Exercises

3.2.1. Verify the Cauchy–Schwarz inequality for each of the following pairs of vectors  $\mathbf{v}, \mathbf{w}$ , using the standard dot product, and then determine the angle between them:

- (a)  $(1, 2)^T, (-1, 2)^T$ , (b)  $(1, -1, 0)^T, (-1, 0, 1)^T$ , (c)  $(1, -1, 0)^T, (2, 2, 2)^T$ ,  
 (d)  $(1, -1, 1, 0)^T, (-2, 0, -1, 1)^T$ , (e)  $(2, 1, -2, -1)^T, (0, -1, 2, -1)^T$ .

3.2.2. (a) Find the Euclidean angle between the vectors  $(1, 1, 1, 1)^T$  and  $(1, 1, 1, -1)^T$  in  $\mathbb{R}^4$ .

(b) List the possible angles between  $(1, 1, 1, 1)^T$  and  $(a_1, a_2, a_3, a_4)^T$ , where each  $a_i$  is either 1 or  $-1$ .

3.2.3. Prove that the points  $(0, 0, 0), (1, 1, 0), (1, 0, 1), (0, 1, 1)$  form the vertices of a regular tetrahedron, meaning that all sides have the same length. What is the common Euclidean angle between the edges? What is the angle between any two rays going from the center  $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$  to the vertices? **Remark.** Methane molecules assume this geometric configuration, and the angle influences their chemistry.

3.2.4. Verify the Cauchy–Schwarz inequality for the vectors  $\mathbf{v} = (1, 2)^T, \mathbf{w} = (1, -3)^T$ , using

- (a) the dot product; (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2$ ;  
 (c) the inner product (3.9).

3.2.5. Verify the Cauchy–Schwarz inequality for the vectors  $\mathbf{v} = (3, -1, 2)^T, \mathbf{w} = (1, -1, 1)^T$ ,

using (a) the dot product; (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$ ;

- (c) the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \mathbf{w}$ .

◊ 3.2.6. Show that one can determine the angle  $\theta$  between  $\mathbf{v}$  and  $\mathbf{w}$  via the formula

$$\cos \theta = \frac{\|\mathbf{v} + \mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2}{4 \|\mathbf{v}\| \|\mathbf{w}\|}.$$

Draw a picture illustrating what is being measured.

◊ 3.2.7. *The Law of Cosines:* Prove that the formula

$$\|\mathbf{v} - \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2 \|\mathbf{v}\| \|\mathbf{w}\| \cos \theta, \quad (3.21)$$

where  $\theta$  is the angle between  $\mathbf{v}$  and  $\mathbf{w}$ , is valid in every inner product space.

- 3.2.8. Use the Cauchy–Schwarz inequality to prove  $(a \cos \theta + b \sin \theta)^2 \leq a^2 + b^2$  for any  $\theta, a, b$ .
- 3.2.9. Prove that  $(a_1 + a_2 + \cdots + a_n)^2 \leq n(a_1^2 + a_2^2 + \cdots + a_n^2)$  for any real numbers  $a_1, \dots, a_n$ . When does equality hold?
- ◇ 3.2.10. The *cross product* of two vectors in  $\mathbb{R}^2$  is defined as the scalar
- $$\mathbf{v} \times \mathbf{w} = v_1 w_2 - v_2 w_1 \quad \text{for} \quad \mathbf{v} = (v_1, v_2)^T, \quad \mathbf{w} = (w_1, w_2)^T. \quad (3.22)$$
- (a) Does the cross product define an inner product on  $\mathbb{R}^2$ ? Carefully explain which axioms are valid and which are not. (b) Prove that  $\mathbf{v} \times \mathbf{w} = \|\mathbf{v}\| \|\mathbf{w}\| \sin \theta$ , where  $\theta$  denotes the angle from  $\mathbf{v}$  to  $\mathbf{w}$  as in Figure 3.2. (c) Prove that  $\mathbf{v} \times \mathbf{w} = 0$  if and only if  $\mathbf{v}$  and  $\mathbf{w}$  are parallel vectors. (d) Show that  $|\mathbf{v} \times \mathbf{w}|$  equals the area of the parallelogram defined by  $\mathbf{v}$  and  $\mathbf{w}$ .
- ◇ 3.2.11. Explain why the inequality  $\langle \mathbf{v}, \mathbf{w} \rangle \leq \|\mathbf{v}\| \|\mathbf{w}\|$ , obtained by omitting the absolute value sign on the left-hand side of Cauchy–Schwarz, is valid.
- 3.2.12. Verify the Cauchy–Schwarz inequality for the functions  $f(x) = x$  and  $g(x) = e^x$  with respect to (a) the  $L^2$  inner product on the interval  $[0, 1]$ , (b) the  $L^2$  inner product on  $[-1, 1]$ , (c) the weighted inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) e^{-x} dx$ .
- 3.2.13. Using the  $L^2$  inner product on the interval  $[0, \pi]$ , find the angle between the functions (a) 1 and  $\cos x$ ; (b) 1 and  $\sin x$ ; (c)  $\cos x$  and  $\sin x$ .
- 3.2.14. Verify the Cauchy–Schwarz inequality for the two particular functions appearing in Exercise 3.1.32 using the  $L^2$  inner product on (a) the unit square; (b) the unit disk.

## Orthogonal Vectors

In Euclidean geometry, a particularly noteworthy configuration occurs when two vectors are *perpendicular*. Perpendicular vectors meet at a right angle,  $\theta = \frac{1}{2}\pi$  or  $\frac{3}{2}\pi$ , with  $\cos \theta = 0$ . The angle formula (3.17) implies that the vectors  $\mathbf{v}, \mathbf{w}$  are perpendicular if and only if their dot product vanishes:  $\mathbf{v} \cdot \mathbf{w} = 0$ . Perpendicularity is of interest in general inner product spaces, but, for historical reasons, has been given a more suggestive name.

**Definition 3.6.** Two elements  $\mathbf{v}, \mathbf{w} \in V$  of an inner product space  $V$  are called *orthogonal* if their inner product vanishes:  $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ .

In particular, the zero element is orthogonal to all other vectors:  $\langle \mathbf{0}, \mathbf{v} \rangle = 0$  for all  $\mathbf{v} \in V$ . Orthogonality is a remarkably powerful tool that appears throughout the manifold applications of linear algebra, and often serves to dramatically simplify many computations. We will devote all of Chapter 4 to a detailed exploration of its manifold implications.

**Example 3.7.** The vectors  $\mathbf{v} = (1, 2)^T$  and  $\mathbf{w} = (6, -3)^T$  are orthogonal with respect to the Euclidean dot product in  $\mathbb{R}^2$ , since  $\mathbf{v} \cdot \mathbf{w} = 1 \cdot 6 + 2 \cdot (-3) = 0$ . We deduce that they meet at a right angle. However, these vectors are *not* orthogonal with respect to the weighted inner product (3.8):

$$\langle \mathbf{v}, \mathbf{w} \rangle = \left\langle \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 6 \\ -3 \end{pmatrix} \right\rangle = 2 \cdot 1 \cdot 6 + 5 \cdot 2 \cdot (-3) = -18 \neq 0.$$

Thus, the property of orthogonality, like angles in general, depends upon which inner product is being used.

**Example 3.8.** The polynomials  $p(x) = x$  and  $q(x) = x^2 - \frac{1}{2}$  are orthogonal with respect to the inner product  $\langle p, q \rangle = \int_0^1 p(x) q(x) dx$  on the interval  $[0, 1]$ , since

$$\langle x, x^2 - \frac{1}{2} \rangle = \int_0^1 x \left( x^2 - \frac{1}{2} \right) dx = \int_0^1 \left( x^3 - \frac{1}{2} x \right) dx = 0.$$

They fail to be orthogonal on most other intervals. For example, on the interval  $[0, 2]$ ,

$$\langle x, x^2 - \frac{1}{2} \rangle = \int_0^2 x \left( x^2 - \frac{1}{2} \right) dx = \int_0^2 \left( x^3 - \frac{1}{2} x \right) dx = 3.$$

**Warning.** There is no obvious connection between the orthogonality of two functions and the geometry of their graphs.

## Exercises

*Note:* Unless stated otherwise, the inner product is the standard dot product on  $\mathbb{R}^n$ .

3.2.15. (a) Find  $a$  such that  $(2, a, -3)^T$  is orthogonal to  $(-1, 3, -2)^T$ . (b) Is there any value of  $a$  for which  $(2, a, -3)^T$  is parallel to  $(-1, 3, -2)^T$ ?

3.2.16. Find all vectors in  $\mathbb{R}^3$  that are orthogonal to both  $(1, 2, 3)^T$  and  $(-2, 0, 1)^T$ .

3.2.17. Answer Exercises 3.2.15 and 3.2.16 using the weighted inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle = 3v_1 w_1 + 2v_2 w_2 + v_3 w_3.$$

3.2.18. Find all vectors in  $\mathbb{R}^4$  that are orthogonal to both  $(1, 2, 3, 4)^T$  and  $(5, 6, 7, 8)^T$ .

3.2.19. Determine a basis for the subspace  $W \subset \mathbb{R}^4$  consisting of all vectors which are orthogonal to the vector  $(1, 2, -1, 3)^T$ .

3.2.20. Find three vectors  $\mathbf{u}, \mathbf{v}$  and  $\mathbf{w}$  in  $\mathbb{R}^3$  such that  $\mathbf{u}$  and  $\mathbf{v}$  are orthogonal,  $\mathbf{u}$  and  $\mathbf{w}$  are orthogonal, but  $\mathbf{v}$  and  $\mathbf{w}$  are *not* orthogonal. Are your vectors linearly independent or linearly dependent? Can you find vectors of the opposite dependency satisfying the same conditions? Why or why not?

3.2.21. For what values of  $a, b$  are the vectors  $(1, 1, a)^T$  and  $(b, -1, 1)^T$  orthogonal

(a) with respect to the dot product?

(b) with respect to the weighted inner product of Exercise 3.2.17?

3.2.22. When is a vector orthogonal to itself?

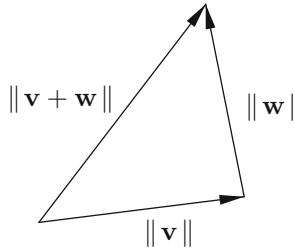
◇ 3.2.23. Prove that the only element  $\mathbf{w}$  in an inner product space  $V$  that is orthogonal to every vector, so  $\langle \mathbf{w}, \mathbf{v} \rangle = 0$  for all  $\mathbf{v} \in V$ , is the zero vector:  $\mathbf{w} = \mathbf{0}$ .

3.2.24. A vector with  $\|\mathbf{v}\| = 1$  is known as a *unit vector*. Prove that if  $\mathbf{v}, \mathbf{w}$  are both unit vectors, then  $\mathbf{v} + \mathbf{w}$  and  $\mathbf{v} - \mathbf{w}$  are orthogonal. Are they also unit vectors?

◇ 3.2.25. Let  $V$  be an inner product space and  $\mathbf{v} \in V$ . Prove that the set of all vectors  $\mathbf{w} \in V$  that are orthogonal to  $\mathbf{v}$  is a subspace of  $V$ .

3.2.26. (a) Show that the polynomials  $p_1(x) = 1$ ,  $p_2(x) = x - \frac{1}{2}$ ,  $p_3(x) = x^2 - x + \frac{1}{6}$  are mutually orthogonal with respect to the  $L^2$  inner product on the interval  $[0, 1]$ .

(b) Show that the functions  $\sin n\pi x$ ,  $n = 1, 2, 3, \dots$ , are mutually orthogonal with respect to the same inner product.



**Figure 3.3.** Triangle Inequality.

- 3.2.27. Find a non-zero quadratic polynomial that is orthogonal to both  $p_1(x) = 1$  and  $p_2(x) = x$  under the  $L^2$  inner product on the interval  $[-1, 1]$ .
- 3.2.28. Find all quadratic polynomials that are orthogonal to the function  $e^x$  with respect to the  $L^2$  inner product on the interval  $[0, 1]$ .
- 3.2.29. Determine all pairs among the functions  $1, x, \cos \pi x, \sin \pi x, e^x$ , that are orthogonal with respect to the  $L^2$  inner product on  $[-1, 1]$ .
- 3.2.30. Find two non-zero functions that are orthogonal with respect to the weighted inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) x dx$ .

### The Triangle Inequality

The familiar triangle inequality states that the length of one side of a triangle is at most equal to the sum of the lengths of the other two sides. Referring to Figure 3.3, if the first two sides are represented by vectors  $\mathbf{v}$  and  $\mathbf{w}$ , then the third corresponds to their sum  $\mathbf{v} + \mathbf{w}$ . The triangle inequality turns out to be an elementary consequence of the Cauchy–Schwarz inequality (3.18), and hence is valid in *every* inner product space.

**Theorem 3.9.** The norm associated with an inner product satisfies the *triangle inequality*

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\| \quad \text{for all } \mathbf{v}, \mathbf{w} \in V. \quad (3.23)$$

Equality holds if and only if  $\mathbf{v}$  and  $\mathbf{w}$  are parallel vectors.

*Proof:* We compute

$$\begin{aligned} \|\mathbf{v} + \mathbf{w}\|^2 &= \langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle = \|\mathbf{v}\|^2 + 2 \langle \mathbf{v}, \mathbf{w} \rangle + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2 \|\mathbf{v}\| \|\mathbf{w}\| + \|\mathbf{w}\|^2 = (\|\mathbf{v}\| + \|\mathbf{w}\|)^2, \end{aligned}$$

where the middle inequality follows from Cauchy–Schwarz, cf. Exercise 3.2.11. Taking square roots of both sides and using the fact that the resulting expressions are both positive completes the proof. *Q.E.D.*

**Example 3.10.** The vectors  $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$  and  $\mathbf{w} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}$  sum to  $\mathbf{v} + \mathbf{w} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}$ . Their

Euclidean norms are  $\|\mathbf{v}\| = \sqrt{6}$  and  $\|\mathbf{w}\| = \sqrt{13}$ , while  $\|\mathbf{v} + \mathbf{w}\| = \sqrt{17}$ . The triangle inequality (3.23) in this case says  $\sqrt{17} \leq \sqrt{6} + \sqrt{13}$ , which is true.

**Example 3.11.** Consider the functions  $f(x) = x - 1$  and  $g(x) = x^2 + 1$ . Using the  $L^2$  norm on the interval  $[0, 1]$ , we find that

$$\begin{aligned}\|f\| &= \sqrt{\int_0^1 (x-1)^2 dx} = \sqrt{\frac{1}{3}}, & \|g\| &= \sqrt{\int_0^1 (x^2+1)^2 dx} = \sqrt{\frac{28}{15}}, \\ \|f+g\| &= \sqrt{\int_0^1 (x^2+x)^2 dx} = \sqrt{\frac{31}{30}}.\end{aligned}$$

The triangle inequality requires  $\sqrt{\frac{31}{30}} \leq \sqrt{\frac{1}{3}} + \sqrt{\frac{28}{15}}$ , which is valid.

The Cauchy–Schwarz and triangle inequalities look much more impressive when written out in full detail. For the Euclidean dot product (3.1), they are

$$\begin{aligned}\left| \sum_{i=1}^n v_i w_i \right| &\leq \sqrt{\sum_{i=1}^n v_i^2} \sqrt{\sum_{i=1}^n w_i^2}, \\ \sqrt{\sum_{i=1}^n (v_i + w_i)^2} &\leq \sqrt{\sum_{i=1}^n v_i^2} + \sqrt{\sum_{i=1}^n w_i^2}.\end{aligned}\tag{3.24}$$

Theorems 3.5 and 3.9 imply that these inequalities are valid for arbitrary real numbers  $v_1, \dots, v_n, w_1, \dots, w_n$ . For the  $L^2$  inner product (3.13) on function space, they produce the following splendid integral inequalities:

$$\begin{aligned}\left| \int_a^b f(x) g(x) dx \right| &\leq \sqrt{\int_a^b f(x)^2 dx} \sqrt{\int_a^b g(x)^2 dx}, \\ \sqrt{\int_a^b [f(x) + g(x)]^2 dx} &\leq \sqrt{\int_a^b f(x)^2 dx} + \sqrt{\int_a^b g(x)^2 dx},\end{aligned}\tag{3.25}$$

which hold for arbitrary continuous (and, in fact, rather general) functions. The first of these is the original Cauchy–Schwarz inequality, whose proof appeared to be quite deep when it first appeared. Only after the abstract notion of an inner product space was properly formalized did its innate simplicity and generality become evident.

## Exercises

- 3.2.31. Use the dot product on  $\mathbb{R}^3$  to answer the following: (a) Find the angle between the vectors  $(1, 2, 3)^T$  and  $(1, -1, 2)^T$ . (b) Verify the Cauchy–Schwarz and triangle inequalities for these two particular vectors. (c) Find all vectors that are orthogonal to both of these vectors.
- 3.2.32. Verify the triangle inequality for each pair of vectors in Exercise 3.2.1.
- 3.2.33. Verify the triangle inequality for the vectors and inner products in Exercise 3.2.4.
- 3.2.34. Verify the triangle inequality for the functions in Exercise 3.2.12 for the indicated inner products.

3.2.35. Verify the triangle inequality for the two particular functions appearing in Exercise 3.1.32 with respect to the  $L^2$  inner product on (a) the unit square; (b) the unit disk.

3.2.36. Use the  $L^2$  inner product  $\langle f, g \rangle = \int_{-1}^1 f(x) g(x) dx$  to answer the following:

(a) Find the “angle” between the functions  $1$  and  $x$ . Are they orthogonal? (b) Verify the Cauchy–Schwarz and triangle inequalities for these two functions. (c) Find all quadratic polynomials  $p(x) = a + bx + cx^2$  that are orthogonal to both of these functions.

3.2.37. (a) Write down the explicit formulae for the Cauchy–Schwarz and triangle inequalities based on the weighted inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) e^x dx$ . (b) Verify that the inequalities hold when  $f(x) = 1$ ,  $g(x) = e^x$  by direct computation. (c) What is the “angle” between these two functions in this inner product?

3.2.38. Answer Exercise 3.2.37 for the Sobolev  $H^1$  inner product

$$\langle f, g \rangle = \int_0^1 [f(x)g(x) + f'(x)g'(x)] dx, \text{ cf. Exercise 3.1.27.}$$

3.2.39. Prove that  $\|\mathbf{v} - \mathbf{w}\| \geq |\|\mathbf{v}\| - \|\mathbf{w}\||$ . Interpret this result pictorially.

3.2.40. *True or false:*  $\|\mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{v} + \mathbf{w}\|$  for all  $\mathbf{v}, \mathbf{w} \in V$ .

♡ 3.2.41. (a) Prove that the space  $\mathbb{R}^\infty$  consisting of all infinite sequences  $\mathbf{x} = (x_1, x_2, x_3, \dots)$  of real numbers  $x_i \in \mathbb{R}$  is a vector space. (b) Prove that the set of all sequences  $\mathbf{x}$  such that  $\sum_{k=1}^{\infty} x_k^2 < \infty$  is a subspace, commonly denoted by  $\ell^2 \subset \mathbb{R}^\infty$ . (c) Write down two examples of sequences  $\mathbf{x}$  belonging to  $\ell^2$  and two that do not belong to  $\ell^2$ . (d) *True or false:* If  $\mathbf{x} \in \ell^2$ , then  $x_k \rightarrow 0$  and  $k \rightarrow \infty$ . (e) *True or false:* If  $x_k \rightarrow 0$  as  $k \rightarrow \infty$ , then  $\mathbf{x} \in \ell^2$ . (f) Given  $\alpha \in \mathbb{R}$ , let  $\mathbf{x}$  be the sequence with  $x_k = \alpha^k$ . For which values of  $\alpha$  is  $\mathbf{x} \in \ell^2$ ? (g) Answer part (f) when  $x_k = k^\alpha$ . (h) Prove that  $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^{\infty} x_k y_k$  defines an inner product on the vector space  $\ell^2$ . What is the corresponding norm? (i) Write out the Cauchy–Schwarz and triangle inequalities for the inner product space  $\ell^2$ .

### 3.3 Norms

Every inner product gives rise to a norm that can be used to measure the magnitude or length of the elements of the underlying vector space. However, not every norm that is used in analysis and applications arises from an inner product. To define a general norm on a vector space, we will extract those properties that do not directly rely on the inner product structure.

**Definition 3.12.** A *norm* on a vector space  $V$  assigns a non-negative real number  $\|\mathbf{v}\|$  to each vector  $\mathbf{v} \in V$ , subject to the following axioms, valid for every  $\mathbf{v}, \mathbf{w} \in V$  and  $c \in \mathbb{R}$ :

- (i) *Positivity:*  $\|\mathbf{v}\| \geq 0$ , with  $\|\mathbf{v}\| = 0$  if and only if  $\mathbf{v} = \mathbf{0}$ .
- (ii) *Homogeneity:*  $\|c\mathbf{v}\| = |c| \|\mathbf{v}\|$ .
- (iii) *Triangle inequality:*  $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ .

As we now know, every inner product gives rise to a norm. Indeed, positivity of the norm is one of the inner product axioms. The homogeneity property follows since

$$\|c\mathbf{v}\| = \sqrt{\langle c\mathbf{v}, c\mathbf{v} \rangle} = \sqrt{c^2 \langle \mathbf{v}, \mathbf{v} \rangle} = |c| \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = |c| \|\mathbf{v}\|.$$

Finally, the triangle inequality for an inner product norm was established in Theorem 3.9. Let us introduce some of the principal examples of norms that do not come from inner products.

First, let  $V = \mathbb{R}^n$ . The  $1$  norm of a vector  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T$  is defined as the sum of the absolute values of its entries:

$$\|\mathbf{v}\|_1 = |v_1| + |v_2| + \dots + |v_n|. \quad (3.26)$$

The  $\max$  or  $\infty$  norm is equal to its maximal entry (in absolute value):

$$\|\mathbf{v}\|_\infty = \max \{ |v_1|, |v_2|, \dots, |v_n| \}. \quad (3.27)$$

Verification of the positivity and homogeneity properties for these two norms is straightforward; the triangle inequality is a direct consequence of the elementary inequality

$$|a + b| \leq |a| + |b|, \quad a, b \in \mathbb{R},$$

for absolute values.

The Euclidean norm,  $1$  norm, and  $\infty$  norm on  $\mathbb{R}^n$  are just three representatives of the general  $p$  norm

$$\|\mathbf{v}\|_p = \sqrt[p]{\sum_{i=1}^n |v_i|^p}. \quad (3.28)$$

This quantity defines a norm for all  $1 \leq p < \infty$ . The  $\infty$  norm is a limiting case of (3.28) as  $p \rightarrow \infty$ . Note that the Euclidean norm (3.3) is the  $2$  norm, and is often designated as such; it is the only  $p$  norm which comes from an inner product. The positivity and homogeneity properties of the  $p$  norm are not hard to establish. The triangle inequality, however, is not trivial; in detail, it reads

$$\sqrt[p]{\sum_{i=1}^n |v_i + w_i|^p} \leq \sqrt[p]{\sum_{i=1}^n |v_i|^p} + \sqrt[p]{\sum_{i=1}^n |w_i|^p}, \quad (3.29)$$

and is known as *Minkowski's inequality*. A complete proof can be found in [50].

There are analogous norms on the space  $C^0[a, b]$  of continuous functions on an interval  $[a, b]$ . Basically, one replaces the previous sums by integrals. Thus, the  $L^p$  norm is defined as

$$\|f\|_p = \sqrt[p]{\int_a^b |f(x)|^p dx}. \quad (3.30)$$

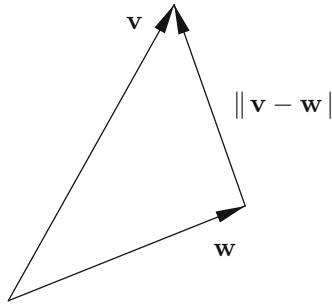
In particular, the  $L^1$  norm is given by integrating the absolute value of the function:

$$\|f\|_1 = \int_a^b |f(x)| dx. \quad (3.31)$$

The  $L^2$  norm (3.13) appears as a special case,  $p = 2$ , and, again, is the only one arising from an inner product. The limiting  $L^\infty$  norm is defined by the maximum

$$\|f\|_\infty = \max \{ |f(x)| : a \leq x \leq b \}. \quad (3.32)$$

Positivity of the  $L^p$  norms again relies on the fact that the only continuous non-negative function with zero integral is the zero function. Homogeneity is easily established. On the



**Figure 3.4.** Distance Between Vectors.

other hand, the proof of the general triangle, or Minkowski, inequality for  $p \neq 1, 2, \infty$  is again not trivial, [19, 68].

**Example 3.13.** Consider the polynomial  $p(x) = 3x^2 - 2$  on the interval  $-1 \leq x \leq 1$ . Its  $L^2$  norm is

$$\|p\|_2 = \sqrt{\int_{-1}^1 (3x^2 - 2)^2 dx} = \sqrt{\frac{18}{5}} = 1.8973\dots$$

Its  $L^\infty$  norm is

$$\|p\|_\infty = \max \{ |3x^2 - 2| : -1 \leq x \leq 1 \} = 2,$$

with the maximum occurring at  $x = 0$ . Finally, its  $L^1$  norm is

$$\begin{aligned} \|p\|_1 &= \int_{-1}^1 |3x^2 - 2| dx \\ &= \int_{-1}^{-\sqrt{2/3}} (3x^2 - 2) dx + \int_{-\sqrt{2/3}}^{\sqrt{2/3}} (2 - 3x^2) dx + \int_{\sqrt{2/3}}^1 (3x^2 - 2) dx \\ &= \left( \frac{4}{3}\sqrt{\frac{2}{3}} - 1 \right) + \frac{8}{3}\sqrt{\frac{2}{3}} + \left( \frac{4}{3}\sqrt{\frac{2}{3}} - 1 \right) = \frac{16}{3}\sqrt{\frac{2}{3}} - 2 = 2.3546\dots \end{aligned}$$

Every norm defines a *distance* between vector space elements, namely

$$d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|. \quad (3.33)$$

For the standard dot product norm, we recover the usual notion of distance between points in Euclidean space. Other types of norms produce alternative (and sometimes quite useful) notions of distance that are, nevertheless, subject to all the familiar properties:

- (a) *Symmetry*:  $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$ ;
- (b) *Positivity*:  $d(\mathbf{v}, \mathbf{w}) = 0$  if and only if  $\mathbf{v} = \mathbf{w}$ ;
- (c) *Triangle inequality*:  $d(\mathbf{v}, \mathbf{w}) \leq d(\mathbf{v}, \mathbf{z}) + d(\mathbf{z}, \mathbf{w})$ .

Just as the distance between vectors measures how close they are to each other — keeping in mind that this measure of proximity depends on the underlying choice of norm — so the distance between functions in a normed function space tells something about how close they are to each other, which is related, albeit subtly, to how close their graphs are. Thus, the norm serves to define the *topology* of the underlying vector space, which determines notions of open and closed sets, convergence, and so on, [19, 68].

## Exercises

3.3.1. Compute the 1, 2, 3, and  $\infty$  norms of the vectors  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ . Verify the triangle inequality in each case.

3.3.2. Answer Exercise 3.3.1 for (a)  $\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 \\ -2 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ -3 \end{pmatrix}$ .

3.3.3. Which two of the vectors  $\mathbf{u} = (-2, 2, 1)^T$ ,  $\mathbf{v} = (1, 4, 1)^T$ ,  $\mathbf{w} = (0, 0, -1)^T$  are closest to each other in distance for (a) the Euclidean norm? (b) the  $\infty$  norm? (c) the 1 norm?

3.3.4. (a) Compute the  $L^\infty$  norm on  $[0, 1]$  of the functions  $f(x) = \frac{1}{3} - x$  and  $g(x) = x - x^2$ .  
 (b) Verify the triangle inequality for these two particular functions.

3.3.5. Answer Exercise 3.3.4 using the  $L^1$  norm.

3.3.6. Which two of the functions  $f(x) = 1$ ,  $g(x) = x$ ,  $h(x) = \sin \pi x$  are closest to each other on the interval  $[0, 1]$  under (a) the  $L^1$  norm? (b) the  $L^2$  norm? (c) the  $L^\infty$  norm?

3.3.7. Consider the functions  $f(x) = 1$  and  $g(x) = x - \frac{3}{4}$  as elements of the vector space  $C^0[0, 1]$ . For each of the following norms, compute  $\|f\|$ ,  $\|g\|$ ,  $\|f + g\|$ , and verify the triangle inequality: (a) the  $L^1$  norm; (b) the  $L^2$  norm; (c) the  $L^3$  norm; (d) the  $L^\infty$  norm.

3.3.8. Answer Exercise 3.3.7 when  $f(x) = e^x$  and  $g(x) = e^{-x}$ .

3.3.9. Carefully prove that  $\|(x, y)^T\| = |x| + 2|x - y|$  defines a norm on  $\mathbb{R}^2$ .

3.3.10. Prove that the following formulas define norms on  $\mathbb{R}^2$ : (a)  $\|\mathbf{v}\| = \sqrt{2v_1^2 + 3v_2^2}$ ,  
 (b)  $\|\mathbf{v}\| = \sqrt{2v_1^2 - v_1 v_2 + 2v_2^2}$ , (c)  $\|\mathbf{v}\| = 2|v_1| + |v_2|$ , (d)  $\|\mathbf{v}\| = \max\{2|v_1|, |v_2|\}$ ,  
 (e)  $\|\mathbf{v}\| = \max\{|v_1 - v_2|, |v_1 + v_2|\}$ , (f)  $\|\mathbf{v}\| = |v_1 - v_2| + |v_1 + v_2|$ .

3.3.11. Which of the following formulas define norms on  $\mathbb{R}^3$ ? (a)  $\|\mathbf{v}\| = \sqrt{2v_1^2 + v_2^2 + 3v_3^2}$ ,  
 (b)  $\|\mathbf{v}\| = \sqrt{v_1^2 + 2v_1 v_2 + v_2^2 + v_3^2}$ , (c)  $\|\mathbf{v}\| = \max\{|v_1|, |v_2|, |v_3|\}$ ,  
 (d)  $\|\mathbf{v}\| = |v_1 - v_2| + |v_2 - v_3| + |v_3 - v_1|$ , (e)  $\|\mathbf{v}\| = |v_1| + \max\{|v_2|, |v_3|\}$ .

3.3.12. Prove that two parallel vectors  $\mathbf{v}$  and  $\mathbf{w}$  have the same norm if and only if  $\mathbf{v} = \pm \mathbf{w}$ .

3.3.13. *True or false:* If  $\|\mathbf{v} + \mathbf{w}\| = \|\mathbf{v}\| + \|\mathbf{w}\|$ , then  $\mathbf{v}, \mathbf{w}$  are parallel vectors.

3.3.14. Prove that the  $\infty$  norm on  $\mathbb{R}^2$  does not come from an inner product. *Hint:* Look at Exercise 3.1.13.

3.3.15. Can formula (3.11) be used to define an inner product for (a) the 1 norm  $\|\mathbf{v}\|_1$  on  $\mathbb{R}^2$ ?  
 (b) the  $\infty$  norm  $\|\mathbf{v}\|_\infty$  on  $\mathbb{R}^2$ ?

◇ 3.3.16. Prove that  $\lim_{p \rightarrow \infty} \|\mathbf{v}\|_p = \|\mathbf{v}\|_\infty$  for all  $\mathbf{v} \in \mathbb{R}^2$ .

◇ 3.3.17. Justify the triangle inequality for (a) the  $L^1$  norm (3.31); (b) the  $L^\infty$  norm (3.32).

◇ 3.3.18. Let  $w(x) > 0$  for  $a \leq x \leq b$  be a weight function. (a) Prove that

$\|f\|_{1,w} = \int_a^b |f(x)| w(x) dx$  defines a norm on  $C^0[a, b]$ , called the *weighted L<sup>1</sup> norm*.

(b) Do the same for the *weighted L<sup>∞</sup> norm*  $\|f\|_{\infty,w} = \max\{|f(x)| w(x) : a \leq x \leq b\}$ .

- 3.3.19. Let  $\|\cdot\|_1$  and  $\|\cdot\|_2$  be two different norms on a vector space  $V$ . (a) Prove that  $\|\mathbf{v}\| = \max\{\|\mathbf{v}\|_1, \|\mathbf{v}\|_2\}$  defines a norm on  $V$ . (b) Does  $\|\mathbf{v}\| = \min\{\|\mathbf{v}\|_1, \|\mathbf{v}\|_2\}$  define a norm? (c) Does the arithmetic mean  $\|\mathbf{v}\| = \frac{1}{2}(\|\mathbf{v}\|_1 + \|\mathbf{v}\|_2)$  define a norm? (d) Does the geometric mean  $\|\mathbf{v}\| = \sqrt{\|\mathbf{v}\|_1 \|\mathbf{v}\|_2}$  define a norm?

## Unit Vectors

Let  $V$  be a normed vector space. The elements  $\mathbf{u} \in V$  that have unit norm,  $\|\mathbf{u}\| = 1$ , play a special role, and are known as *unit vectors* (or functions or elements). The following easy lemma shows how to construct a unit vector pointing in the same direction as any given nonzero vector.

**Lemma 3.14.** If  $\mathbf{v} \neq \mathbf{0}$  is any nonzero vector, then the vector  $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|$  obtained by dividing  $\mathbf{v}$  by its norm is a unit vector parallel to  $\mathbf{v}$ .

*Proof:* We compute, making use of the homogeneity property of the norm and the fact that  $\|\mathbf{v}\|$  is a scalar,

$$\|\mathbf{u}\| = \left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} \right\| = \frac{\|\mathbf{v}\|}{\|\mathbf{v}\|} = 1. \quad Q.E.D.$$

**Example 3.15.** The vector  $\mathbf{v} = (1, -2)^T$  has length  $\|\mathbf{v}\|_2 = \sqrt{5}$  with respect to the standard Euclidean norm. Therefore, the unit vector pointing in the same direction is

$$\mathbf{u} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix}.$$

On the other hand, for the 1 norm,  $\|\mathbf{v}\|_1 = 3$ , and so

$$\tilde{\mathbf{u}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_1} = \frac{1}{3} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$

is the unit vector parallel to  $\mathbf{v}$  in the 1 norm. Finally,  $\|\mathbf{v}\|_\infty = 2$ , and hence the corresponding unit vector for the  $\infty$  norm is

$$\hat{\mathbf{u}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_\infty} = \frac{1}{2} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -1 \end{pmatrix}.$$

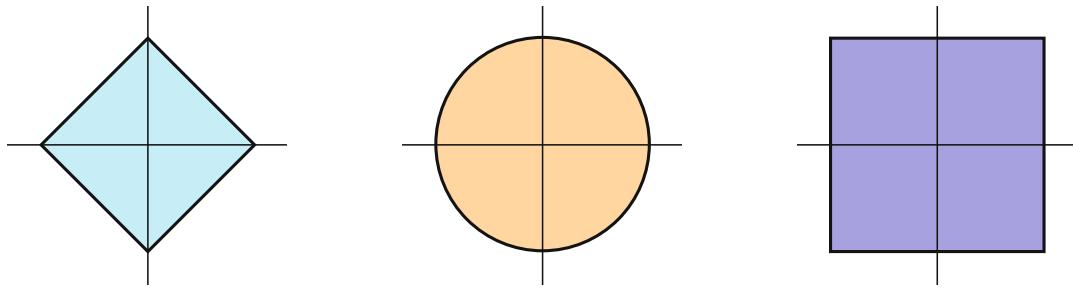
Thus, the notion of unit vector will depend upon which norm is being used.

**Example 3.16.** Similarly, on the interval  $[0, 1]$ , the quadratic polynomial  $p(x) = x^2 - \frac{1}{2}$  has  $L^2$  norm

$$\|p\|_2 = \sqrt{\int_0^1 (x^2 - \frac{1}{2})^2 dx} = \sqrt{\int_0^1 (x^4 - x^2 + \frac{1}{4}) dx} = \sqrt{\frac{7}{60}}.$$

Therefore,  $u(x) = \frac{p(x)}{\|p\|} = \sqrt{\frac{60}{7}} x^2 - \sqrt{\frac{15}{7}}$  is a “unit polynomial”,  $\|u\|_2 = 1$ , which is “parallel” to (or, more precisely, a scalar multiple of) the polynomial  $p$ . On the other hand, for the  $L^\infty$  norm,

$$\|p\|_\infty = \max \left\{ |x^2 - \frac{1}{2}| \mid 0 \leq x \leq 1 \right\} = \frac{1}{2},$$



**Figure 3.5.** Unit Balls and Spheres for 1, 2, and  $\infty$  Norms in  $\mathbb{R}^2$ .

and hence, in this case,  $\tilde{u}(x) = 2p(x) = 2x^2 - 1$  is the corresponding unit polynomial.

The *unit sphere* for the given norm is defined as the set of all unit vectors

$$S_1 = \{ \|\mathbf{u}\| = 1 \}, \quad \text{while} \quad S_r = \{ \|\mathbf{u}\| = r \} \quad (3.34)$$

is the sphere of radius  $r \geq 0$ . Thus, the unit sphere for the Euclidean norm on  $\mathbb{R}^n$  is the usual round sphere

$$S_1 = \{ \|\mathbf{x}\|^2 = x_1^2 + x_2^2 + \dots + x_n^2 = 1 \}.$$

The unit sphere for the  $\infty$  norm is the surface of a unit cube:

$$S_1 = \left\{ \mathbf{x} \in \mathbb{R}^n \mid \begin{array}{l} |x_i| \leq 1, \quad i = 1, \dots, n, \quad \text{and either} \\ x_1 = \pm 1 \text{ or } x_2 = \pm 1 \text{ or } \dots \text{ or } x_n = \pm 1 \end{array} \right\}.$$

For the 1 norm,

$$S_1 = \{ \mathbf{x} \in \mathbb{R}^n \mid |x_1| + |x_2| + \dots + |x_n| = 1 \}$$

is the unit diamond in two dimensions, unit octahedron in three dimensions, and unit *cross polytope* in general. See [Figure 3.5](#) for the two-dimensional pictures.

In all cases, the *unit ball*  $B_1 = \{ \|\mathbf{u}\| \leq 1 \}$  consists of all vectors of norm less than or equal to 1, and has the unit sphere as its boundary. If  $V$  is a finite-dimensional normed vector space, then the unit ball  $B_1$  is a *compact* subset, meaning that it is closed and bounded. This basic topological fact, which is *not* true in infinite-dimensional normed spaces, underscores the distinction between finite-dimensional vector analysis and the vastly more complicated infinite-dimensional realm.

## Exercises

- 3.3.20. Find a unit vector in the same direction as  $\mathbf{v} = (1, 2, -3)^T$  for (a) the Euclidean norm, (b) the weighted norm  $\|\mathbf{v}\|^2 = 2v_1^2 + v_2^2 + \frac{1}{3}v_3^2$ , (c) the 1 norm, (d) the  $\infty$  norm, (e) the norm based on the inner product  $2v_1w_1 - v_1w_2 - v_2w_1 + 2v_2w_2 - v_2w_3 - v_3w_2 + 2v_3w_3$ .
- 3.3.21. Show that, for every choice of given angles  $\theta$ ,  $\phi$ , and  $\psi$ , the following are unit vectors in the Euclidean norm: (a)  $(\cos \theta \cos \phi, \cos \theta \sin \phi, \sin \theta)^T$ . (b)  $\frac{1}{\sqrt{2}}(\cos \theta, \sin \theta, \cos \phi, \sin \phi)^T$ . (c)  $(\cos \theta \cos \phi \cos \psi, \cos \theta \cos \phi \sin \psi, \cos \theta \sin \phi, \sin \theta)^T$ .
- 3.3.22. How many unit vectors are parallel to a given vector  $\mathbf{v} \neq \mathbf{0}$ ? (a) 1, (b) 2, (c) 3, (d)  $\infty$ , (e) depends on the norm. Explain your answer.
- 3.3.23. Plot the unit circle (sphere) for (a) the weighted norm  $\|\mathbf{v}\| = \sqrt{v_1^2 + 4v_2^2}$ ; (b) the norm based on the inner product (3.9); (c) the norm of Exercise 3.3.9.

- 3.3.24. Draw the unit circle for each norm in Exercise 3.3.10.
- 3.3.25. Sketch the unit sphere  $S_1 \subset \mathbb{R}^3$  for (a) the  $L^1$  norm, (b) the  $L^\infty$  norm, (c) the weighted norm  $\|\mathbf{v}\|^2 = 2v_1^2 + v_2^2 + 3v_3^2$ , (d)  $\|\mathbf{v}\| = \max\{|v_1 + v_2|, |v_1 + v_3|, |v_2 + v_3|\}$ .
- 3.3.26. Let  $\mathbf{v} \neq \mathbf{0}$  be any nonzero vector in a normed vector space  $V$ . Show how to construct a new norm on  $V$  that changes  $\mathbf{v}$  into a unit vector.
- 3.3.27. *True or false:* Two norms on a vector space have the same unit sphere if and only if they are the same norm.
- 3.3.28. Find the unit function that is a constant multiple of the function  $f(x) = x - \frac{1}{3}$  with respect to the (a)  $L^1$  norm on  $[0, 1]$ ; (b)  $L^2$  norm on  $[0, 1]$ ; (c)  $L^\infty$  norm on  $[0, 1]$ ; (d)  $L^1$  norm on  $[-1, 1]$ ; (e)  $L^2$  norm on  $[-1, 1]$ ; (f)  $L^\infty$  norm on  $[-1, 1]$ .
- 3.3.29. For which norms is the constant function  $f(x) \equiv 1$  a unit function?  
 (a)  $L^1$  norm on  $[0, 1]$ ; (b)  $L^2$  norm on  $[0, 1]$ ; (c)  $L^\infty$  norm on  $[0, 1]$ ;  
 (d)  $L^1$  norm on  $[-1, 1]$ ; (e)  $L^2$  norm on  $[-1, 1]$ ; (f)  $L^\infty$  norm on  $[-1, 1]$ ;  
 (g)  $L^1$  norm on  $\mathbb{R}$ ; (h)  $L^2$  norm on  $\mathbb{R}$ ; (i)  $L^\infty$  norm on  $\mathbb{R}$ .
- ◊ 3.3.30. A subset  $S \subset \mathbb{R}^n$  is called *convex* if, for all  $\mathbf{x}, \mathbf{y} \in S$ , the line segment joining  $\mathbf{x}$  to  $\mathbf{y}$  is also in  $S$ , i.e.,  $t\mathbf{x} + (1 - t)\mathbf{y} \in S$  for all  $0 \leq t \leq 1$ . Prove that the unit ball is a convex subset of a normed vector space. Is the unit sphere convex?

## Equivalence of Norms

While there are many different types of norms, in a finite-dimensional vector space they are all more or less equivalent. “Equivalence” does not mean that they assume the same values, but rather that they are, in a certain sense, always close to one another, and so, for many analytical purposes, may be used interchangeably. As a consequence, we may be able to simplify the analysis of a problem by choosing a suitably adapted norm; examples can be found in Chapter 9.

**Theorem 3.17.** Let  $\|\cdot\|_1$  and  $\|\cdot\|_2$  be any two norms on  $\mathbb{R}^n$ . Then there exist positive constants  $0 < c^* \leq C^*$  such that

$$c^* \|\mathbf{v}\|_1 \leq \|\mathbf{v}\|_2 \leq C^* \|\mathbf{v}\|_1 \quad \text{for every } \mathbf{v} \in \mathbb{R}^n. \quad (3.35)$$

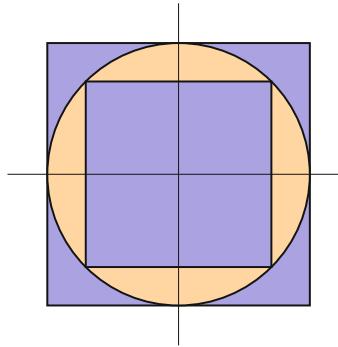
*Proof:* We just sketch the basic idea, leaving the details to a more rigorous real analysis course, cf. [19; §7.6]. We begin by noting that a norm defines a continuous real-valued function  $f(\mathbf{v}) = \|\mathbf{v}\|$  on  $\mathbb{R}^n$ . (Continuity is, in fact, a consequence of the triangle inequality.) Let  $S_1 = \{\|\mathbf{u}\|_1 = 1\}$  denote the unit sphere of the first norm. Every continuous function defined on a compact set achieves both a maximum and a minimum value. Thus, restricting the second norm function to the unit sphere  $S_1$  of the first norm, we can set

$$c^* = \min \{ \|\mathbf{u}\|_2 \mid \mathbf{u} \in S_1 \}, \quad C^* = \max \{ \|\mathbf{u}\|_2 \mid \mathbf{u} \in S_1 \}. \quad (3.36)$$

Moreover,  $0 < c^* \leq C^* < \infty$ , with equality holding if and only if the norms are the same. The minimum and maximum (3.36) will serve as the constants in the desired inequalities (3.35). Indeed, by definition,

$$c^* \leq \|\mathbf{u}\|_2 \leq C^* \quad \text{when } \|\mathbf{u}\|_1 = 1, \quad (3.37)$$

which proves that (3.35) is valid for all unit vectors  $\mathbf{v} = \mathbf{u} \in S_1$ . To prove the inequalities in general, assume  $\mathbf{v} \neq \mathbf{0}$ . (The case  $\mathbf{v} = \mathbf{0}$  is trivial.) Lemma 3.14 says that  $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|_1 \in S_1$



**Figure 3.6.** Equivalence of the  $\infty$  and 2 Norms.

is a unit vector in the first norm:  $\|\mathbf{u}\|_1 = 1$ . Moreover, by the homogeneity property of the norm,  $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2/\|\mathbf{v}\|_1$ . Substituting into (3.37) and clearing denominators completes the proof of (3.35). *Q.E.D.*

**Example 3.18.** Consider the Euclidean norm  $\|\cdot\|_2$  and the max norm  $\|\cdot\|_\infty$  on  $\mathbb{R}^n$ . According to (3.36), the bounding constants are found by minimizing and maximizing  $\|\mathbf{u}\|_\infty = \max\{|u_1|, \dots, |u_n|\}$  over all unit vectors  $\|\mathbf{u}\|_2 = 1$  on the (round) unit sphere. The maximal value is achieved at the poles  $\pm \mathbf{e}_k$ , with  $\|\pm \mathbf{e}_k\|_\infty = C^* = 1$ . The minimal value is attained at the points  $(\pm \frac{1}{\sqrt{n}}, \dots, \pm \frac{1}{\sqrt{n}})$ , whereby  $c^* = \frac{1}{\sqrt{n}}$ . Therefore,

$$\frac{1}{\sqrt{n}} \|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2. \quad (3.38)$$

We can interpret these inequalities as follows. Suppose  $\mathbf{v}$  is a vector lying on the unit sphere in the Euclidean norm, so  $\|\mathbf{v}\|_2 = 1$ . Then (3.38) tells us that its  $\infty$  norm is bounded from above and below by  $\frac{1}{\sqrt{n}} \leq \|\mathbf{v}\|_\infty \leq 1$ . Therefore, the Euclidean unit sphere sits inside the  $\infty$  norm unit sphere and outside the  $\infty$  norm sphere of radius  $\frac{1}{\sqrt{n}}$ . [Figure 3.6](#) illustrates the two-dimensional situation: the unit circle is inside the unit square, and contains the square of size  $\frac{1}{\sqrt{2}}$ .

One significant consequence of the equivalence of norms is that, in  $\mathbb{R}^n$ , convergence is independent of the norm. The following are all equivalent to the standard notion of convergence of a sequence  $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}, \dots$  of vectors in  $\mathbb{R}^n$ :

- (a) the vectors converge:  $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^*$ ;
- (b) the individual coordinates all converge:  $u_i^{(k)} \rightarrow u_i^*$  for  $i = 1, \dots, n$ ;
- (c) the difference in norms goes to zero:  $\|\mathbf{u}^{(k)} - \mathbf{u}^*\| \rightarrow 0$ .

The last version, known as *convergence in norm*, does not depend on which norm is chosen. Indeed, the inequality (3.35) implies that if one norm goes to zero, so does any other norm. A consequence is that all norms on  $\mathbb{R}^n$  induce the same topology — convergence of sequences, notions of open and closed sets, and so on. None of this is true in infinite-dimensional function space! A rigorous development of the underlying topological and analytical properties of compactness, continuity, and convergence is beyond the scope of this course. The motivated student is encouraged to consult a text in real analysis, e.g., [19, 68], to find the relevant definitions, theorems, and proofs.

**Example 3.19.** Consider the infinite-dimensional vector space  $C^0[0, 1]$  consisting of all continuous functions on the interval  $[0, 1]$ . The functions

$$f_n(x) = \begin{cases} 1 - nx, & 0 \leq x \leq \frac{1}{n}, \\ 0, & \frac{1}{n} \leq x \leq 1, \end{cases}$$

have identical  $L^\infty$  norms

$$\|f_n\|_\infty = \sup \{ |f_n(x)| \mid 0 \leq x \leq 1 \} = 1.$$

On the other hand, their  $L^2$  norm

$$\|f_n\|_2 = \sqrt{\int_0^1 f_n(x)^2 dx} = \sqrt{\int_0^{1/n} (1 - nx)^2 dx} = \frac{1}{\sqrt{3n}}$$

goes to zero as  $n \rightarrow \infty$ . This example shows that there is no constant  $C^*$  such that

$$\|f\|_\infty \leq C^* \|f\|_2$$

for all  $f \in C^0[0, 1]$ . Thus, the  $L^\infty$  and  $L^2$  norms on  $C^0[0, 1]$  are not equivalent — there exist functions that have unit  $L^\infty$  norm, but arbitrarily small  $L^2$  norm. Similar comparative results can be established for the other function space norms. Analysis and topology on function space is intimately linked to the underlying choice of norm.

## Exercises

3.3.31. Check the validity of the inequalities (3.38) for the particular vectors

$$(a) (1, -1)^T, \quad (b) (1, 2, 3)^T, \quad (c) (1, 1, 1, 1)^T, \quad (d) (1, -1, -2, -1, 1)^T.$$

3.3.32. Find all  $\mathbf{v} \in \mathbb{R}^2$  such that

$$(a) \|\mathbf{v}\|_1 = \|\mathbf{v}\|_\infty, \quad (b) \|\mathbf{v}\|_1 = \|\mathbf{v}\|_2, \quad (c) \|\mathbf{v}\|_2 = \|\mathbf{v}\|_\infty, \quad (d) \|\mathbf{v}\|_\infty = \frac{1}{\sqrt{2}} \|\mathbf{v}\|_2.$$

3.3.33. How would you quantify the following statement: The norm of a vector is small if and only if all its entries are small.

3.3.34. Can you find an elementary proof of the inequalities  $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2 \leq \sqrt{n} \|\mathbf{v}\|_\infty$  for  $\mathbf{v} \in \mathbb{R}^n$  directly from the formulas for the norms?

3.3.35. (i) Show the equivalence of the Euclidean norm and the 1 norm on  $\mathbb{R}^n$  by proving  $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1 \leq \sqrt{n} \|\mathbf{v}\|_2$ . (ii) Verify that the vectors in Exercise 3.3.31 satisfy both inequalities. (iii) For which vectors  $\mathbf{v} \in \mathbb{R}^n$  is (a)  $\|\mathbf{v}\|_2 = \|\mathbf{v}\|_1$ ? (b)  $\|\mathbf{v}\|_1 = \sqrt{n} \|\mathbf{v}\|_2$ ?

3.3.36. (i) Establish the equivalence inequalities (3.35) between the 1 and  $\infty$  norms.

(ii) Verify them for the vectors in Exercise 3.3.31.

(iii) For which vectors  $\mathbf{v} \in \mathbb{R}^n$  are your inequalities equality?

3.3.37. Let  $\|\cdot\|_2$  denote the usual Euclidean norm on  $\mathbb{R}^n$ . Determine the constants in the norm equivalence inequalities  $c^* \|\mathbf{v}\| \leq \|\mathbf{v}\|_2 \leq C^* \|\mathbf{v}\|$  for the following norms: (a) the weighted norm  $\|\mathbf{v}\| = \sqrt{2v_1^2 + 3v_2^2}$ , (b) the norm  $\|\mathbf{v}\| = \max\{|v_1 + v_2|, |v_1 - v_2|\}$ .

3.3.38. Let  $\|\cdot\|$  be a norm on  $\mathbb{R}^n$ . Prove that there is a constant  $C > 0$  such that the entries of every  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T \in \mathbb{R}^n$  are all bounded, in absolute value, by  $|v_i| \leq C \|\mathbf{v}\|$ .

3.3.39. Prove that if  $[a, b]$  is a bounded interval and  $f \in C^0[a, b]$ , then  $\|f\|_2 \leq \sqrt{b-a} \|f\|_\infty$ .

◇ 3.3.40. In this exercise, the indicated function norms are taken over all of  $\mathbb{R}$ .

(a) Let  $f_n(x) = \begin{cases} 1, & -n \leq x \leq n, \\ 0, & \text{otherwise.} \end{cases}$  Prove that  $\|f_n\|_\infty = 1$ , but  $\|f_n\|_2 \rightarrow \infty$  as  $n \rightarrow \infty$ .

(b) Explain why there is no constant  $C$  such that  $\|f\|_2 \leq C\|f\|_\infty$  for all functions  $f$ .

(c) Let  $f_n(x) = \begin{cases} \sqrt{\frac{n}{2}}, & -\frac{1}{n} \leq x \leq \frac{1}{n}, \\ 0, & \text{otherwise.} \end{cases}$  Prove that  $\|f_n\|_2 = 1$ , but  $\|f_n\|_\infty \rightarrow \infty$

as  $n \rightarrow \infty$ . Conclude that there is no constant  $C$  such that  $\|f\|_\infty \leq C\|f\|_2$ .

(d) Construct similar examples that disprove the related inequalities

(i)  $\|f\|_\infty \leq C\|f\|_1$ , (ii)  $\|f\|_1 \leq C\|f\|_2$ , (iii)  $\|f\|_2 \leq C\|f\|_1$ .

◇ 3.3.41. (a) Prove that the  $L^\infty$  and  $L^2$  norms on the vector space  $C^0[-1, 1]$  are not equivalent.

*Hint:* Look at Exercise 3.3.40 for ideas. (b) Can you establish a bound in either direction, i.e.,  $\|f\|_\infty \leq C\|f\|_2$  or  $\|f\|_2 \leq \tilde{C}\|f\|_\infty$  for all  $f \in C^0[-1, 1]$  for some positive constants  $C, \tilde{C}$ ? (c) Are the  $L^1$  and  $L^\infty$  norms equivalent?

◇ 3.3.42. What does it mean if the constants defined in (3.36) are equal:  $c^* = C^*$ ?

3.3.43. Suppose  $\langle \mathbf{v}, \mathbf{w} \rangle_1$  and  $\langle \mathbf{v}, \mathbf{w} \rangle_2$  are two inner products on the same vector space  $V$ . For which  $\alpha, \beta \in \mathbb{R}$  is the linear combination  $\langle \mathbf{v}, \mathbf{w} \rangle = \alpha \langle \mathbf{v}, \mathbf{w} \rangle_1 + \beta \langle \mathbf{v}, \mathbf{w} \rangle_2$  a legitimate inner product? *Hint:* The case  $\alpha, \beta \geq 0$  is easy. However, some negative values are also permitted, and your task is to decide which.

◇ 3.3.44. Suppose  $\|\cdot\|_1, \|\cdot\|_2$  are two norms on  $\mathbb{R}^n$ . Prove that the corresponding matrix norms satisfy  $\hat{c}^* \|A\|_1 \leq \|A\|_2 \leq \tilde{C}^* \|A\|_1$  for any  $n \times n$  matrix  $A$  for some positive constants  $0 < \hat{c}^* \leq \tilde{C}^*$ .

## Matrix Norms

Each norm on  $\mathbb{R}^n$  will naturally induce a norm on the vector space  $\mathcal{M}_{n \times n}$  of all  $n \times n$  matrices. Roughly speaking, the matrix norm tells us how much a linear transformation stretches vectors relative to the given norm. Matrix norms will play an important role in Chapters 8 and 9, particularly in our analysis of linear iterative systems and iterative numerical methods for solving both linear and nonlinear systems.

We work exclusively with real  $n \times n$  matrices in this section, although the results straightforwardly extend to complex matrices. We begin by fixing a norm  $\|\cdot\|$  on  $\mathbb{R}^n$ . The norm may or may not come from an inner product — this is irrelevant as far as the construction goes.

**Theorem 3.20.** If  $\|\cdot\|$  is any norm on  $\mathbb{R}^n$ , then the quantity

$$\|A\| = \max \{ \|A\mathbf{u}\| \mid \|\mathbf{u}\| = 1 \} \quad (3.39)$$

defines the norm of an  $n \times n$  matrix  $A \in \mathcal{M}_{n \times n}$ , called the associated *natural matrix norm*.

*Proof:* First note that  $\|A\| < \infty$ , since the maximum is taken on a closed and bounded subset, namely the unit sphere  $S_1 = \{\|\mathbf{u}\| = 1\}$  for the given norm. To show that (3.39) defines a norm, we need to verify the three basic axioms of Definition 3.12.

Non-negativity,  $\|A\| \geq 0$ , is immediate. Suppose  $\|A\| = 0$ . This means that, for every unit vector,  $\|A\mathbf{u}\| = 0$ , and hence  $A\mathbf{u} = \mathbf{0}$  whenever  $\|\mathbf{u}\| = 1$ . If  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$  is any nonzero vector, then  $\mathbf{u} = \mathbf{v}/r$ , where  $r = \|\mathbf{v}\|$ , is a unit vector, so

$$A\mathbf{v} = A(r\mathbf{u}) = rA\mathbf{u} = \mathbf{0}. \quad (3.40)$$

Therefore,  $A\mathbf{v} = \mathbf{0}$  for every  $\mathbf{v} \in \mathbb{R}^n$ , which implies that  $A = \mathbf{O}$  is the zero matrix. This serves to prove the positivity property:  $\|A\| = 0$  if and only if  $A = \mathbf{O}$ .

As for homogeneity, if  $c \in \mathbb{R}$  is any scalar, then

$$\|cA\| = \max \{ \|cA\mathbf{u}\| \} = \max \{ |c| \|A\mathbf{u}\| \} = |c| \max \{ \|A\mathbf{u}\| \} = |c| \|A\|.$$

Finally, to prove the triangle inequality, we use the fact that the maximum of the sum of quantities is bounded by the sum of their individual maxima. Therefore, since the norm on  $\mathbb{R}^n$  satisfies the triangle inequality,

$$\begin{aligned} \|A + B\| &= \max \{ \|A\mathbf{u} + B\mathbf{u}\| \} \leq \max \{ \|A\mathbf{u}\| + \|B\mathbf{u}\| \} \\ &\leq \max \{ \|A\mathbf{u}\| \} + \max \{ \|B\mathbf{u}\| \} = \|A\| + \|B\|. \end{aligned} \quad Q.E.D.$$

The property that distinguishes a matrix norm from a generic norm on the space of matrices is the fact that it also obeys a very useful *product inequality*.

**Theorem 3.21.** A natural matrix norm satisfies

$$\|A\mathbf{v}\| \leq \|A\| \|\mathbf{v}\|, \quad \text{for all } A \in \mathcal{M}_{n \times n}, \quad \mathbf{v} \in \mathbb{R}^n. \quad (3.41)$$

Furthermore,

$$\|AB\| \leq \|A\| \|B\|, \quad \text{for all } A, B \in \mathcal{M}_{n \times n}. \quad (3.42)$$

*Proof:* Note first that, by definition  $\|A\mathbf{u}\| \leq \|A\|$  for all unit vectors  $\|\mathbf{u}\| = 1$ . Then, letting  $\mathbf{v} = r\mathbf{u}$  where  $\mathbf{u}$  is a unit vector and  $r = \|\mathbf{v}\|$ , we have

$$\|A\mathbf{v}\| = \|A(r\mathbf{u})\| = r\|A\mathbf{u}\| \leq r\|A\| = \|\mathbf{v}\| \|A\|,$$

proving the first inequality. To prove the second, we apply the first, replacing  $\mathbf{v}$  by  $B\mathbf{u}$ :

$$\begin{aligned} \|AB\| &= \max \{ \|AB\mathbf{u}\| \} = \max \{ \|A(B\mathbf{u})\| \} \\ &\leq \max \{ \|A\| \|B\mathbf{u}\| \} = \|A\| \max \{ \|B\mathbf{u}\| \} = \|A\| \|B\|. \end{aligned} \quad Q.E.D.$$

**Remark.** In general, a norm on the vector space of  $n \times n$  matrices is called a *matrix norm* if it also satisfies the multiplicative inequality (3.42). Most, but not all, matrix norms used in applications come from norms on the underlying vector space.

The multiplicative inequality (3.42) implies, in particular, that  $\|A^2\| \leq \|A\|^2$ ; equality is not necessarily valid. More generally:

**Proposition 3.22.** If  $A$  is a square matrix, then  $\|A^k\| \leq \|A\|^k$ .

Let us determine the explicit formula for the matrix norm induced by the  $\infty$  norm

$$\|\mathbf{v}\|_\infty = \max \{ |v_1|, \dots, |v_n| \}.$$

The corresponding formula for the 1 norm is left as Exercise 3.3.48. The formula for the Euclidean matrix norm (2 norm) will be deferred until Theorem 8.71.

**Definition 3.23.** The  $i^{\text{th}}$  *absolute row sum* of a matrix  $A$  is the sum of the absolute values of the entries in the  $i^{\text{th}}$  row:

$$s_i = |a_{i1}| + \dots + |a_{in}| = \sum_{j=1}^n |a_{ij}|. \quad (3.43)$$

**Proposition 3.24.** The  $\infty$  *matrix norm* of a matrix  $A$  is equal to its maximal absolute row sum:

$$\|A\|_\infty = \max\{s_1, \dots, s_n\} = \max \left\{ \sum_{j=1}^n |a_{ij}| \mid 1 \leq i \leq n \right\}. \quad (3.44)$$

*Proof:* Let  $s = \max\{s_1, \dots, s_n\}$  denote the right-hand side of (3.44). Given any  $\mathbf{v} \in \mathbb{R}^n$ , we compute the  $\infty$  norm of the image vector  $A\mathbf{v}$ :

$$\begin{aligned} \|A\mathbf{v}\|_\infty &= \max \left\{ \left| \sum_{j=1}^n a_{ij}v_j \right| \right\} \leq \max \left\{ \sum_{j=1}^n |a_{ij}v_j| \right\} \\ &\leq \max \left\{ \sum_{j=1}^n |a_{ij}| \right\} \max \{ |v_j| \} = s \|\mathbf{v}\|_\infty. \end{aligned}$$

In particular, by specializing to a unit vector,  $\|\mathbf{v}\|_\infty = 1$ , we deduce that  $\|A\|_\infty \leq s$ .

On the other hand, suppose the maximal absolute row sum occurs at row  $i$ , so

$$s_i = \sum_{j=1}^n |a_{ij}| = s. \quad (3.45)$$

Let  $\mathbf{u} \in \mathbb{R}^n$  be the specific vector that has the following entries:  $u_j = +1$  if  $a_{ij} \geq 0$ , while  $u_j = -1$  if  $a_{ij} < 0$ . Then  $\|\mathbf{u}\|_\infty = 1$ . Moreover, since  $a_{ij}u_j = |a_{ij}|$ , the  $i^{\text{th}}$  entry of  $A\mathbf{u}$  is equal to the  $i^{\text{th}}$  absolute row sum (3.45). This implies that  $\|A\|_\infty \geq \|A\mathbf{u}\|_\infty \geq s$ . *Q.E.D.*

**Example 3.25.** Consider the symmetric matrix  $A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{4} \end{pmatrix}$ . Its two absolute row sums are  $\left| \frac{1}{2} \right| + \left| -\frac{1}{3} \right| = \frac{5}{6}$ ,  $\left| -\frac{1}{3} \right| + \left| \frac{1}{4} \right| = \frac{7}{12}$ , so  $\|A\|_\infty = \max \{ \frac{5}{6}, \frac{7}{12} \} = \frac{5}{6}$ .

## Exercises

3.3.45. Compute the  $\infty$  matrix norm of the following matrices.

$$(a) \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{6} \end{pmatrix}, \quad (b) \begin{pmatrix} \frac{5}{3} & \frac{4}{3} \\ -\frac{7}{6} & -\frac{5}{6} \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & .1 & .8 \\ -.1 & 0 & .1 \\ -.8 & -.1 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{pmatrix}.$$

3.3.46. Find a matrix  $A$  such that  $\|A^2\|_\infty \neq \|A\|_\infty^2$ .

3.3.47. *True or false:* If  $B = S^{-1}AS$  are similar matrices, then  $\|B\|_\infty = \|A\|_\infty$ .

◇ 3.3.48. (i) Find an explicit formula for the 1 matrix norm  $\|A\|_1$ .

(ii) Compute the 1 matrix norm of the matrices in Exercise 3.3.45.

3.3.49. Prove directly from the axioms of Definition 3.12 that (3.44) defines a norm on the space of  $n \times n$  matrices.

3.3.50. Let  $A = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$ . Compute the natural matrix norm  $\|A\|$  for (a) the weighted  $\infty$  norm  $\|\mathbf{v}\| = \max\{2|v_1|, 3|v_2|\}$ ; (b) the weighted 1 norm  $\|\mathbf{v}\| = 2|v_1| + 3|v_2|$ .

3.3.51. The *Frobenius norm* of an  $n \times n$  matrix  $A$  is defined as  $\|A\|_F = \sqrt{\sum_{i,j=1}^n a_{ij}^2}$ .

Prove that this defines a matrix norm by checking the three norm axioms plus the multiplicative inequality (3.42).

3.3.52. Explain why  $\|A\| = \max|a_{ij}|$  defines a norm on the space of  $n \times n$  matrices. Show by example that this is *not* a matrix norm, i.e., (3.42) is not necessarily valid.

## 3.4 Positive Definite Matrices

Let us now return to the study of inner products and fix our attention on the finite-dimensional situation. Our immediate goal is to determine the most general inner product that can be placed on the finite-dimensional vector space  $\mathbb{R}^n$ . The answer will lead us to the important class of positive definite matrices, which appear in a wide range of applications, including minimization problems, mechanics, electrical circuits, differential equations, statistics, and numerical methods. Moreover, their infinite-dimensional counterparts, positive definite linear operators, govern most boundary value problems arising in continuum physics and engineering.

Suppose we are given an inner product  $\langle \mathbf{x}, \mathbf{y} \rangle$  between vectors  $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_n)^T$  and  $\mathbf{y} = (y_1 \ y_2 \ \dots \ y_n)^T$  in  $\mathbb{R}^n$ . Our goal is to determine its explicit formula. We begin by writing the vectors in terms of the standard basis vectors (2.17):

$$\mathbf{x} = x_1 \mathbf{e}_1 + \dots + x_n \mathbf{e}_n = \sum_{i=1}^n x_i \mathbf{e}_i, \quad \mathbf{y} = y_1 \mathbf{e}_1 + \dots + y_n \mathbf{e}_n = \sum_{j=1}^n y_j \mathbf{e}_j. \quad (3.46)$$

To evaluate their inner product, we will appeal to the three basic axioms. We first employ bilinearity to expand

$$\langle \mathbf{x}, \mathbf{y} \rangle = \left\langle \sum_{i=1}^n x_i \mathbf{e}_i, \sum_{j=1}^n y_j \mathbf{e}_j \right\rangle = \sum_{i,j=1}^n x_i y_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle.$$

Therefore,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i,j=1}^n k_{ij} x_i y_j = \mathbf{x}^T K \mathbf{y}, \quad (3.47)$$

where  $K$  denotes the  $n \times n$  matrix of inner products of the basis vectors, with entries

$$k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle, \quad i, j = 1, \dots, n. \quad (3.48)$$

We conclude that any inner product must be expressed in the general *bilinear form* (3.47).

The two remaining inner product axioms will impose certain constraints on the inner product matrix  $K$ . Symmetry implies that

$$k_{ij} = \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \langle \mathbf{e}_j, \mathbf{e}_i \rangle = k_{ji}, \quad i, j = 1, \dots, n.$$

Consequently, the inner product matrix  $K$  must be symmetric:

$$K = K^T.$$

Conversely, symmetry of  $K$  ensures symmetry of the bilinear form:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y} = (\mathbf{x}^T K \mathbf{y})^T = \mathbf{y}^T K^T \mathbf{x} = \mathbf{y}^T K \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle,$$

where the second equality follows from the fact that the quantity  $\mathbf{x}^T K \mathbf{y}$  is a scalar, and hence equals its transpose.

The final condition for an inner product is positivity, which requires that

$$\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \quad (3.49)$$

with equality if and only if  $\mathbf{x} = \mathbf{0}$ . The precise meaning of this positivity condition on the matrix  $K$  is not so immediately evident, and so will be encapsulated in a definition.

**Definition 3.26.** An  $n \times n$  matrix  $K$  is called *positive definite* if it is symmetric,  $K^T = K$ , and satisfies the positivity condition

$$\mathbf{x}^T K \mathbf{x} > 0 \quad \text{for all } \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n. \quad (3.50)$$

We will sometimes write  $K > 0$  to mean that  $K$  is a positive definite matrix.

**Warning.** The condition  $K > 0$  does *not* mean that all the entries of  $K$  are positive. There are many positive definite matrices that have some negative entries; see Example 3.28 below. Conversely, many symmetric matrices with all positive entries are not positive definite!

**Remark.** Although some authors allow non-symmetric matrices to be designated as positive definite, we will say that a matrix is positive definite *only* when it is symmetric. But, to underscore our convention and remind the casual reader, we will often include the superfluous adjective “symmetric” when speaking of positive definite matrices.

Our preliminary analysis has resulted in the following general characterization of inner products on a finite-dimensional vector space.

**Theorem 3.27.** Every inner product on  $\mathbb{R}^n$  is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y} \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \quad (3.51)$$

where  $K$  is a symmetric, positive definite  $n \times n$  matrix.

Given a symmetric matrix  $K$ , the homogeneous quadratic polynomial

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j, \quad (3.52)$$

is known as a *quadratic form*<sup>†</sup> on  $\mathbb{R}^n$ . The quadratic form is called *positive definite* if

$$q(\mathbf{x}) > 0 \quad \text{for all } \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n. \quad (3.53)$$

So the quadratic form (3.52) is positive definite if and only if its coefficient matrix  $K$  is.

---

<sup>†</sup> Exercise 3.4.15 shows that the coefficient matrix  $K$  in any quadratic form can be taken to be symmetric without any loss of generality.

**Example 3.28.** Even though the symmetric matrix  $K = \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix}$  has two negative entries, it is, nevertheless, a positive definite matrix. Indeed, the corresponding quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = 4x_1^2 - 4x_1x_2 + 3x_2^2 = (2x_1 - x_2)^2 + 2x_2^2 \geq 0$$

is a sum of two non-negative quantities. Moreover,  $q(\mathbf{x}) = 0$  if and only if both  $2x_1 - x_2 = 0$  and  $x_2 = 0$ , which implies  $x_1 = 0$  also. This proves  $q(\mathbf{x}) > 0$  for all  $\mathbf{x} \neq \mathbf{0}$ , and hence  $K$  is indeed a positive definite matrix. The corresponding inner product on  $\mathbb{R}^2$  is

$$\langle \mathbf{x}, \mathbf{y} \rangle = (x_1 \ x_2) \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 4x_1y_1 - 2x_1y_2 - 2x_2y_1 + 3x_2y_2.$$

On the other hand, despite the fact that  $K = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$  has all positive entries, it is *not* a positive definite matrix. Indeed, writing out

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = x_1^2 + 4x_1x_2 + x_2^2,$$

we find, for instance, that  $q(1, -1) = -2 < 0$ , violating positivity. These two simple examples should be enough to convince the reader that the problem of determining whether a given symmetric matrix is positive definite is not completely elementary.

**Example 3.29.** By definition, a general symmetric  $2 \times 2$  matrix  $K = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$  is positive definite if and only if the associated quadratic form satisfies

$$q(\mathbf{x}) = ax_1^2 + 2bx_1x_2 + cx_2^2 > 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}. \quad (3.54)$$

Analytic geometry tells us that this is the case if and only if

$$a > 0, \quad ac - b^2 > 0, \quad (3.55)$$

i.e., the quadratic form has positive leading coefficient and positive determinant (or negative discriminant). A direct proof of this well-known fact will appear shortly.

With a little practice, it is not difficult to read off the coefficient matrix  $K$  from the explicit formula for the quadratic form (3.52).

**Example 3.30.** Consider the quadratic form

$$q(x, y, z) = x^2 + 4xy + 6y^2 - 2xz + 9z^2$$

depending upon three variables. The corresponding coefficient matrix is

$$K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix} \quad \text{whereby} \quad q(x, y, z) = (x \ y \ z) \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Note that the squared terms in  $q$  contribute directly to the diagonal entries of  $K$ , while the mixed terms are split in half to give the symmetric off-diagonal entries. As a challenge, the reader might wish to try proving that this particular matrix is positive definite by establishing positivity of the quadratic form:  $q(x, y, z) > 0$  for all nonzero  $(x, y, z)^T \in \mathbb{R}^3$ . Later, we will devise a simple, systematic test for positive definiteness.

Slightly more generally, a quadratic form and its associated symmetric coefficient matrix are called *positive semi-definite* if

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \quad (3.56)$$

in which case we write  $K \geq 0$ . A positive semi-definite matrix may have *null directions*, meaning non-zero vectors  $\mathbf{z}$  such that  $q(\mathbf{z}) = \mathbf{z}^T K \mathbf{z} = 0$ . Clearly, every nonzero vector  $\mathbf{z} \in \ker K$  that lies in the coefficient matrix's kernel defines a null direction, but there may be others. A positive definite matrix is not allowed to have null directions, and so  $\ker K = \{\mathbf{0}\}$ . Recalling Proposition 2.42, we deduce that all positive definite matrices are nonsingular. The converse, however, is *not* valid; many symmetric, nonsingular matrices fail to be positive definite.

**Proposition 3.31.** If a matrix is positive definite, then it is nonsingular.

**Example 3.32.** The matrix  $K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$  is positive semi-definite, but not positive definite. Indeed, the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = x_1^2 - 2x_1 x_2 + x_2^2 = (x_1 - x_2)^2 \geq 0$$

is a perfect square, and so clearly non-negative. However, the elements of  $\ker K$ , namely the scalar multiples of the vector  $(1, 1)^T$ , define null directions:  $q(c, c) = 0$ .

In a similar fashion, a quadratic form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$  and its associated symmetric matrix  $K$  are called *negative semi-definite* if  $q(\mathbf{x}) \leq 0$  for all  $\mathbf{x}$  and *negative definite* if  $q(\mathbf{x}) < 0$  for all  $\mathbf{x} \neq \mathbf{0}$ . A quadratic form is called *indefinite* if it is neither positive nor negative semi-definite, equivalently, if there exist points  $\mathbf{x}_+$  where  $q(\mathbf{x}_+) > 0$  and points  $\mathbf{x}_-$  where  $q(\mathbf{x}_-) < 0$ . Details can be found in the exercises.

Only positive definite matrices define inner products. However, indefinite matrices play a fundamental role in Einstein's theory of special relativity, [55]. In particular, the quadratic form associated with the matrix

$$K = \begin{pmatrix} c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{namely} \quad q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = c^2 t^2 - x^2 - y^2 - z^2 \quad \text{where} \quad \mathbf{x} = \begin{pmatrix} t \\ x \\ y \\ z \end{pmatrix}, \quad (3.57)$$

with  $c$  representing the speed of light, is the so-called Minkowski "metric" on relativistic space-time  $\mathbb{R}^4$ . The null directions form the light cone; see Exercise 3.4.20.

## Exercises

3.4.1. Which of the following  $2 \times 2$  matrices are positive definite?

- (a)  $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 5 & 3 \\ 3 & -2 \end{pmatrix}$ , (e)  $\begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}$ .

In the positive definite cases, write down the formula for the associated inner product.

3.4.2. Let  $K = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$ . Prove that the associated quadratic form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$  is indefinite by finding a point  $\mathbf{x}^+$  where  $q(\mathbf{x}^+) > 0$  and a point  $\mathbf{x}^-$  where  $q(\mathbf{x}^-) < 0$ .

◇ 3.4.3. (a) Prove that a diagonal matrix  $D = \text{diag}(c_1, c_2, \dots, c_n)$  is positive definite if and only if all its diagonal entries are positive:  $c_i > 0$ . (b) Write down and identify the associated inner product.

3.4.4. Write out the Cauchy–Schwarz and triangle inequalities for the inner product defined in Example 3.28.

- ◇ 3.4.5. (a) Show that every diagonal entry of a positive definite matrix must be positive.  
 (b) Write down a symmetric matrix with all positive diagonal entries that is not positive definite. (c) Find a nonzero matrix with one or more zero diagonal entries that is positive semi-definite.
- 3.4.6. Prove that if  $K$  is any positive definite matrix, then every positive scalar multiple  $cK$ ,  $c > 0$ , is also positive definite.
- ◇ 3.4.7. (a) Show that if  $K$  and  $L$  are positive definite matrices, so is  $K + L$ . (b) Give an example of two matrices that are not positive definite whose sum is positive definite.
- 3.4.8. Find two positive definite matrices  $K$  and  $L$  whose product  $KL$  is not positive definite.
- 3.4.9. Write down a nonsingular symmetric matrix that is not positive or negative definite.
- ◇ 3.4.10. Let  $K$  be a nonsingular symmetric matrix. (a) Show that  $\mathbf{x}^T K^{-1} \mathbf{x} = \mathbf{y}^T K \mathbf{y}$ , where  $K \mathbf{y} = \mathbf{x}$ . (b) Prove that if  $K$  is positive definite, then so is  $K^{-1}$ .
- ◇ 3.4.11. Prove that an  $n \times n$  symmetric matrix  $K$  is positive definite if and only if, for every  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$ , the vectors  $\mathbf{v}$  and  $K\mathbf{v}$  meet at an acute Euclidean angle:  $|\theta| < \frac{1}{2}\pi$ .
- ◇ 3.4.12. Prove that the inner product associated with a positive definite quadratic form  $q(\mathbf{x})$  is given by the *polarization formula*  $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2} [q(\mathbf{x} + \mathbf{y}) - q(\mathbf{x}) - q(\mathbf{y})]$ .
- 3.4.13. (a) Is it possible for a quadratic form to be positive,  $q(\mathbf{x}_+) > 0$ , at only one point  $\mathbf{x}_+ \in \mathbb{R}^n$ ? (b) Under what conditions is  $q(\mathbf{x}_0) = 0$  at only one point?
- ◇ 3.4.14. (a) Let  $K$  and  $L$  be symmetric  $n \times n$  matrices. Prove that  $\mathbf{x}^T K \mathbf{x} = \mathbf{x}^T L \mathbf{x}$  for all  $\mathbf{x} \in \mathbb{R}^n$  if and only if  $K = L$ . (b) Find an example of two non-symmetric matrices  $K \neq L$  such that  $\mathbf{x}^T K \mathbf{x} = \mathbf{x}^T L \mathbf{x}$  for all  $\mathbf{x} \in \mathbb{R}^n$ .
- ◇ 3.4.15. Suppose  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} = \sum_{i,j=1}^n a_{ij} x_i x_j$  is a general quadratic form on  $\mathbb{R}^n$ , whose coefficient matrix  $A$  is not necessarily symmetric. Prove that  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$ , where  $K = \frac{1}{2}(A + A^T)$  is a symmetric matrix. Therefore, we do not lose any generality by restricting our discussion to quadratic forms that are constructed from symmetric matrices.
- 3.4.16. (a) Show that a symmetric matrix  $N$  is negative definite if and only if  $K = -N$  is positive definite. (b) Write down two explicit criteria that tell whether a  $2 \times 2$  matrix  $N = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$  is negative definite. (c) Use your criteria to check whether  
 (i)  $\begin{pmatrix} -1 & 1 \\ 1 & -2 \end{pmatrix}$ , (ii)  $\begin{pmatrix} -4 & -5 \\ -5 & -6 \end{pmatrix}$ , (iii)  $\begin{pmatrix} -3 & -1 \\ -1 & 2 \end{pmatrix}$  are negative definite.
- 3.4.17. Show that  $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  is a null direction for  $K = \begin{pmatrix} 1 & -2 \\ -2 & 3 \end{pmatrix}$ , but  $\mathbf{x} \notin \ker K$ .
- 3.4.18. Explain why an indefinite quadratic form necessarily has a non-zero null direction.
- 3.4.19. Let  $K = K^T$ . *True or false:* (a) If  $K$  admits a null direction, then  $\ker K \neq \{\mathbf{0}\}$ .  
 (b) If  $K$  has no null directions, then  $K$  is either positive or negative definite.
- ◇ 3.4.20. In special relativity, light rays in Minkowski space-time  $\mathbb{R}^n$  travel along the *light cone* which, by definition, consists of all null directions associated with an indefinite quadratic form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$ . Find and sketch a picture of the light cone when the coefficient matrix  $K$  is (a)  $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$ . **Remark.** In the physical universe, space-time is  $n = 4$ -dimensional, and  $K$  is given in (3.57), [55].

- ◇ 3.4.21. A function  $f(\mathbf{x})$  on  $\mathbb{R}^n$  is called *homogeneous* of degree  $k$  if  $f(c\mathbf{x}) = c^k f(\mathbf{x})$  for all scalars  $c$ . (a) Given  $\mathbf{a} \in \mathbb{R}^n$ , show that the *linear form*  $\ell(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} = a_1 x_1 + \cdots + a_n x_n$  is homogeneous of degree 1. (b) Show that the quadratic form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j$  is homogeneous of degree 2. (c) Find a homogeneous function of degree 2 on  $\mathbb{R}^2$  that is not a quadratic form.

## Gram Matrices

Symmetric matrices whose entries are given by inner products of elements of an inner product space will appear throughout this text. They are named after the nineteenth-century Danish mathematician Jørgen Gram — not the metric mass unit!

**Definition 3.33.** Let  $V$  be an inner product space, and let  $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ . The associated *Gram matrix*

$$K = \begin{pmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \dots & \langle \mathbf{v}_1, \mathbf{v}_n \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \dots & \langle \mathbf{v}_2, \mathbf{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_n, \mathbf{v}_1 \rangle & \langle \mathbf{v}_n, \mathbf{v}_2 \rangle & \dots & \langle \mathbf{v}_n, \mathbf{v}_n \rangle \end{pmatrix} \quad (3.58)$$

is the  $n \times n$  matrix whose entries are the inner products between the selected vector space elements.

Symmetry of the inner product implies symmetry of the Gram matrix:

$$k_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \langle \mathbf{v}_j, \mathbf{v}_i \rangle = k_{ji}, \quad \text{and hence} \quad K^T = K. \quad (3.59)$$

In fact, the most direct method for producing positive definite and semi-definite matrices is through the Gram matrix construction.

**Theorem 3.34.** All Gram matrices are positive semi-definite. The Gram matrix (3.58) is positive definite if and only if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are linearly independent.

*Proof:* To prove positive (semi-)definiteness of  $K$ , we need to examine the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^n k_{ij} x_i x_j.$$

Substituting the values (3.59) for the matrix entries, we obtain

$$q(\mathbf{x}) = \sum_{i,j=1}^n \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j.$$

Bilinearity of the inner product on  $V$  implies that we can assemble this summation into a single inner product

$$q(\mathbf{x}) = \left\langle \sum_{i=1}^n x_i \mathbf{v}_i, \sum_{j=1}^n x_j \mathbf{v}_j \right\rangle = \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|^2 \geq 0, \quad \text{where} \quad \mathbf{v} = \sum_{i=1}^n x_i \mathbf{v}_i$$

lies in the subspace of  $V$  spanned by the given vectors. This immediately proves that  $K$  is positive semi-definite.

Moreover,  $q(\mathbf{x}) = \|\mathbf{v}\|^2 > 0$  as long as  $\mathbf{v} \neq \mathbf{0}$ . If  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are linearly independent, then

$$\mathbf{v} = x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n = \mathbf{0} \quad \text{if and only if} \quad x_1 = \dots = x_n = 0,$$

and hence  $q(\mathbf{x}) = 0$  if and only if  $\mathbf{x} = \mathbf{0}$ . This implies that  $q(\mathbf{x})$  and hence  $K$  are positive definite. *Q.E.D.*

**Example 3.35.** Consider the vectors  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 6 \end{pmatrix}$ . For the standard Euclidean dot product on  $\mathbb{R}^3$ , the Gram matrix is

$$K = \begin{pmatrix} \mathbf{v}_1 \cdot \mathbf{v}_1 & \mathbf{v}_1 \cdot \mathbf{v}_2 \\ \mathbf{v}_2 \cdot \mathbf{v}_1 & \mathbf{v}_2 \cdot \mathbf{v}_2 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 45 \end{pmatrix}.$$

Since  $\mathbf{v}_1, \mathbf{v}_2$  are linearly independent,  $K > 0$ . Positive definiteness implies that

$$q(x_1, x_2) = 6x_1^2 - 6x_1x_2 + 45x_2^2 > 0 \quad \text{for all } (x_1, x_2) \neq \mathbf{0}.$$

Indeed, this can be checked directly, by using the criteria in (3.55).

On the other hand, for the weighted inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle = 3v_1w_1 + 2v_2w_2 + 5v_3w_3, \quad (3.60)$$

the corresponding Gram matrix is

$$\tilde{K} = \begin{pmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \end{pmatrix} = \begin{pmatrix} 16 & -21 \\ -21 & 207 \end{pmatrix}. \quad (3.61)$$

Since  $\mathbf{v}_1, \mathbf{v}_2$  are still linearly independent (which, of course, does not depend upon which inner product is used), the matrix  $\tilde{K}$  is also positive definite.

In the case of the Euclidean dot product, the construction of the Gram matrix  $K$  can be directly implemented as follows. Given column vectors  $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^m$ , let us form the  $m \times n$  matrix  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ . In view of the identification (3.2) between the dot product and multiplication of row and column vectors, the  $(i, j)$  entry of  $K$  is given as the product

$$k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_j$$

of the  $i^{\text{th}}$  row of the transpose  $A^T$  and the  $j^{\text{th}}$  column of  $A$ . In other words, the Gram matrix can be evaluated as a matrix product:

$$K = A^T A. \quad (3.62)$$

For the preceding Example 3.35,

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix}, \quad \text{and so} \quad K = A^T A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 6 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 45 \end{pmatrix}.$$

Theorem 3.34 implies that the Gram matrix (3.62) is positive definite if and only if the columns of  $A$  are linearly independent vectors. This implies the following result.

**Proposition 3.36.** Given an  $m \times n$  matrix  $A$ , the following are equivalent:

- (a) The  $n \times n$  Gram matrix  $K = A^T A$  is positive definite.
- (b)  $A$  has linearly independent columns.
- (c)  $\text{rank } A = n \leq m$ .
- (d)  $\ker A = \{0\}$ .

Changing the underlying inner product will, of course, change the Gram matrix. As noted in Theorem 3.27, every inner product on  $\mathbb{R}^m$  has the form

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w} \quad \text{for} \quad \mathbf{v}, \mathbf{w} \in \mathbb{R}^m, \quad (3.63)$$

where  $C > 0$  is a symmetric, positive definite  $m \times m$  matrix. Therefore, given  $n$  vectors  $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^m$ , the entries of the Gram matrix with respect to this inner product are

$$k_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \mathbf{v}_i^T C \mathbf{v}_j.$$

If, as above, we assemble the column vectors into an  $m \times n$  matrix  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ , then the Gram matrix entry  $k_{ij}$  is obtained by multiplying the  $i^{\text{th}}$  row of  $A^T$  by the  $j^{\text{th}}$  column of the product matrix  $CA$ . Therefore, the Gram matrix based on the alternative inner product (3.63) is given by

$$K = A^T C A. \quad (3.64)$$

Theorem 3.34 immediately implies that  $K$  is positive definite — provided that the matrix  $A$  has rank  $n$ .

**Theorem 3.37.** Suppose  $A$  is an  $m \times n$  matrix with linearly independent columns. Suppose  $C$  is any positive definite  $m \times m$  matrix. Then the Gram matrix  $K = A^T C A$  is a positive definite  $n \times n$  matrix.

The Gram matrices constructed in (3.64) arise in a wide variety of applications, including least squares approximation theory (cf. Chapter 5), and mechanical structures and electrical circuits (cf. Chapters 6 and 10). In the majority of applications,  $C = \text{diag}(c_1, \dots, c_m)$  is a diagonal positive definite matrix, which requires it to have strictly positive diagonal entries  $c_i > 0$ . This choice corresponds to a weighted inner product (3.10) on  $\mathbb{R}^m$ .

**Example 3.38.** Returning to the situation of Example 3.35, the weighted inner product

(3.60) corresponds to the diagonal positive definite matrix  $C = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$ . Therefore,

the weighted Gram matrix (3.64) based on the vectors  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 6 \end{pmatrix}$ , is

$$\tilde{K} = A^T C A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 16 & -21 \\ -21 & 207 \end{pmatrix},$$

reproducing (3.61).

The Gram matrix construction is not restricted to finite-dimensional vector spaces, but also applies to inner products on function space. Here is a particularly important example.

**Example 3.39.** Consider the vector space  $C^0[0, 1]$  consisting of continuous functions on the interval  $0 \leq x \leq 1$ , equipped with the  $L^2$  inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) dx$ . Let us construct the Gram matrix corresponding to the simple monomial functions  $1, x, x^2$ .

We compute the required inner products

$$\begin{aligned}\langle 1, 1 \rangle &= \|1\|^2 = \int_0^1 dx = 1, & \langle 1, x \rangle &= \int_0^1 x dx = \frac{1}{2}, \\ \langle x, x \rangle &= \|x\|^2 = \int_0^1 x^2 dx = \frac{1}{3}, & \langle 1, x^2 \rangle &= \int_0^1 x^2 dx = \frac{1}{3}, \\ \langle x^2, x^2 \rangle &= \|x^2\|^2 = \int_0^1 x^4 dx = \frac{1}{5}, & \langle x, x^2 \rangle &= \int_0^1 x^3 dx = \frac{1}{4}.\end{aligned}$$

Therefore, the Gram matrix is

$$K = \begin{pmatrix} \langle 1, 1 \rangle & \langle 1, x \rangle & \langle 1, x^2 \rangle \\ \langle x, 1 \rangle & \langle x, x \rangle & \langle x, x^2 \rangle \\ \langle x^2, 1 \rangle & \langle x^2, x \rangle & \langle x^2, x^2 \rangle \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

As we know, the monomial functions  $1, x, x^2$  are linearly independent, and so Theorem 3.34 immediately implies that the matrix  $K$  is positive definite.

The alert reader may recognize this particular Gram matrix as the  $3 \times 3$  Hilbert matrix that we encountered in (1.72). More generally, the Gram matrix corresponding to the monomials  $1, x, x^2, \dots, x^n$  has entries

$$k_{ij} = \langle x^{i-1}, x^{j-1} \rangle = \int_0^1 x^{i+j-2} dx = \frac{1}{i+j-1}, \quad i, j = 1, \dots, n+1,$$

and is thus the  $(n+1) \times (n+1)$  Hilbert matrix (1.72):  $K = H_{n+1}$ . As a consequence of Theorem 3.34 and Proposition 3.31 (and also Exercise 2.3.36), we have proved the following non-trivial result.

**Proposition 3.40.** The  $n \times n$  Hilbert matrix  $H_n$  is positive definite. Consequently,  $H_n$  is a nonsingular matrix.

**Example 3.41.** Let us construct the Gram matrix corresponding to the trigonometric functions  $1, \cos x, \sin x$ , with respect to the inner product  $\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) g(x) dx$  on the interval  $[-\pi, \pi]$ . We compute the inner products

$$\begin{aligned}\langle 1, 1 \rangle &= \|1\|^2 = \int_{-\pi}^{\pi} dx = 2\pi, & \langle 1, \cos x \rangle &= \int_{-\pi}^{\pi} \cos x dx = 0, \\ \langle \cos x, \cos x \rangle &= \|\cos x\|^2 = \int_{-\pi}^{\pi} \cos^2 x dx = \pi, & \langle 1, \sin x \rangle &= \int_{-\pi}^{\pi} \sin x dx = 0, \\ \langle \sin x, \sin x \rangle &= \|\sin x\|^2 = \int_{-\pi}^{\pi} \sin^2 x dx = \pi, & \langle \cos x, \sin x \rangle &= \int_{-\pi}^{\pi} \cos x \sin x dx = 0.\end{aligned}$$

Therefore, the Gram matrix is a simple diagonal matrix:  $K = \begin{pmatrix} 2\pi & 0 & 0 \\ 0 & \pi & 0 \\ 0 & 0 & \pi \end{pmatrix}$ . Positive definiteness of  $K$  is immediately evident.

If the columns of  $A$  are linearly dependent, then the associated Gram matrix is only positive semi-definite. In this case, the Gram matrix will have nontrivial null directions  $\mathbf{v}$ , so that  $\mathbf{0} \neq \mathbf{v} \in \ker K = \ker A$ .

**Proposition 3.42.** Let  $K = A^T C A$  be the  $n \times n$  Gram matrix constructed from an  $m \times n$  matrix  $A$  and a positive definite  $m \times m$  matrix  $C > 0$ . Then  $\ker K = \ker A$ , and hence  $\text{rank } K = \text{rank } A$ .

*Proof:* Clearly, if  $A\mathbf{x} = \mathbf{0}$ , then  $K\mathbf{x} = A^T C A \mathbf{x} = \mathbf{0}$ , and so  $\ker A \subset \ker K$ . Conversely, if  $K\mathbf{x} = \mathbf{0}$ , then

$$0 = \mathbf{x}^T K \mathbf{x} = \mathbf{x}^T A^T C A \mathbf{x} = \mathbf{y}^T C \mathbf{y}, \quad \text{where } \mathbf{y} = A \mathbf{x}.$$

Since  $C > 0$ , this implies  $\mathbf{y} = \mathbf{0}$ , and hence  $\mathbf{x} \in \ker A$ . Finally, by Theorem 2.49,  $\text{rank } K = n - \dim \ker K = n - \dim \ker A = \text{rank } A$ . *Q.E.D.*

## Exercises

3.4.22. (a) Find the Gram matrix corresponding to each of the following sets of vectors using

the Euclidean dot product on  $\mathbb{R}^n$ . (i)  $\begin{pmatrix} -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ ,  
 (iii)  $\begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ 0 \\ 2 \end{pmatrix}$ , (iv)  $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ , (v)  $\begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}$ ,  
 (vi)  $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \\ -4 \end{pmatrix}, \begin{pmatrix} -1 \\ 3 \\ -1 \end{pmatrix}$ , (vii)  $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -3 \end{pmatrix}$ .

(b) Which are positive definite? (c) If the matrix is positive semi-definite, find all its null directions.

3.4.23. Recompute the Gram matrices for cases (iii)–(v) in the previous exercise using the weighted inner product  $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + 2 x_2 y_2 + 3 x_3 y_3$ . Does this change its positive definiteness?

3.4.24. Recompute the Gram matrices for cases (vi)–(viii) in Exercise 3.4.22 for the weighted inner product  $\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + \frac{1}{2} x_2 y_2 + \frac{1}{3} x_3 y_3 + \frac{1}{4} x_4 y_4$ .

3.4.25. Find the Gram matrix  $K$  for the functions  $1, e^x, e^{2x}$  using the  $L^2$  inner product on  $[0, 1]$ . Is  $K$  positive definite?

3.4.26. Answer Exercise 3.4.25 using the weighted inner product  $\langle f, g \rangle = \int_0^1 f(x) g(x) e^{-x} dx$ .

3.4.27. Find the Gram matrix  $K$  for the monomials  $1, x, x^2, x^3$  using the  $L^2$  inner product on  $[-1, 1]$ . Is  $K$  positive definite?

3.4.28. Answer Exercise 3.4.27 using the weighted inner product  $\langle f, g \rangle = \int_{-1}^1 f(x) g(x) (1 + x) dx$ .

3.4.29. Let  $K$  be a  $2 \times 2$  Gram matrix. Explain why the positive definiteness criterion (3.55) is equivalent to the Cauchy–Schwarz inequality.

◇ 3.4.30. (a) Prove that if  $K$  is a positive definite matrix, then  $K^2$  is also positive definite.

(b) More generally, if  $S = S^T$  is symmetric and nonsingular, then  $S^2$  is positive definite.

3.4.31. Let  $A$  be an  $m \times n$  matrix. (a) Explain why the product  $L = AA^T$  is a Gram matrix.

(b) Show that, even though they may be of different sizes, both Gram matrices  $K = A^T A$  and  $L = AA^T$  have the same rank. (c) Under what conditions are both  $K$  and  $L$  positive definite?

◊ 3.4.32. Let  $K = A^T C A$ , where  $C > 0$ . Prove that

- (a)  $\ker K = \text{coker } K = \ker A$ ; (b)  $\text{img } K = \text{coimg } K = \text{coimg } A$ .

◊ 3.4.33. Prove that every positive definite matrix  $K$  can be written as a Gram matrix.

3.4.34. Suppose  $K$  is the Gram matrix computed from  $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$  relative to a given inner product. Let  $\tilde{K}$  be the Gram matrix for the same elements, but computed relative to a *different* inner product. Show that  $K > 0$  if and only if  $\tilde{K} > 0$ .

◊ 3.4.35. Let  $K_1 = A_1^T C_1 A_1$  and  $K_2 = A_2^T C_2 A_2$  be any two  $n \times n$  Gram matrices. Let  $K = K_1 + K_2$ . (a) Show that if  $K_1, K_2 > 0$  then  $K > 0$ . (b) Give an example in which  $K_1$  and  $K_2$  are not positive definite, but  $K > 0$ . (c) Show that  $K$  is also a Gram matrix, by finding a matrix  $A$  such that  $K = A^T C A$ . *Hint:*  $A$  will have size  $(m_1 + m_2) \times n$ , where  $m_1$  and  $m_2$  are the numbers of rows in  $A_1, A_2$ , respectively.

3.4.36. Show that  $\mathbf{0} \neq \mathbf{z}$  is a null direction for the quadratic form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$  based on the Gram matrix  $K = A^T C A$  if and only if  $\mathbf{z} \in \ker K$ .

## 3.5 Completing the Square

Gram matrices furnish us with an almost inexhaustible supply of positive definite matrices. However, we still do not know how to test whether a given symmetric matrix is positive definite. As we shall soon see, the secret already appears in the particular computations in Examples 3.2 and 3.28.

You may recall the algebraic technique known as “completing the square”, first arising in the derivation of the formula for the solution to the quadratic equation

$$q(x) = ax^2 + 2bx + c = 0, \quad (3.65)$$

and, later, helping to facilitate the integration of various types of rational and algebraic functions. The idea is to combine the first two terms in (3.65) as a perfect square, and thereby rewrite the quadratic function in the form

$$q(x) = a \left( x + \frac{b}{a} \right)^2 + \frac{ac - b^2}{a} = 0. \quad (3.66)$$

As a consequence,

$$\left( x + \frac{b}{a} \right)^2 = \frac{b^2 - ac}{a^2}.$$

The familiar *quadratic formula*

$$x = \frac{-b \pm \sqrt{b^2 - ac}}{a}$$

follows by taking the square root of both sides and then solving for  $x$ . The intermediate step (3.66), where we eliminate the linear term, is known as *completing the square*.

We can perform the same kind of manipulation on a homogeneous quadratic form

$$q(x_1, x_2) = ax_1^2 + 2bx_1 x_2 + cx_2^2. \quad (3.67)$$

In this case, provided  $a \neq 0$ , completing the square amounts to writing

$$q(x_1, x_2) = ax_1^2 + 2bx_1 x_2 + cx_2^2 = a \left( x_1 + \frac{b}{a} x_2 \right)^2 + \frac{ac - b^2}{a} x_2^2 = a y_1^2 + \frac{ac - b^2}{a} y_2^2. \quad (3.68)$$

The net result is to re-express  $q(x_1, x_2)$  as a simpler sum of squares of the new variables

$$y_1 = x_1 + \frac{b}{a} x_2, \quad y_2 = x_2. \quad (3.69)$$

It is not hard to see that the final expression in (3.68) is positive definite, as a function of  $y_1, y_2$ , if and only if both coefficients are positive:

$$a > 0, \quad \frac{ac - b^2}{a} > 0. \quad (3.70)$$

Therefore,  $q(x_1, x_2) \geq 0$ , with equality if and only if  $y_1 = y_2 = 0$ , or, equivalently,  $x_1 = x_2 = 0$ . This conclusively proves that conditions (3.70) are necessary and sufficient for the quadratic form (3.67) to be positive definite.

Our goal is to adapt this simple idea to analyze the positivity of quadratic forms depending on more than two variables. To this end, let us rewrite the quadratic form identity (3.68) in matrix form. The original quadratic form (3.67) is

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}, \quad \text{where} \quad K = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \quad (3.71)$$

Similarly, the right-hand side of (3.68) can be written as

$$\hat{q}(\mathbf{y}) = \mathbf{y}^T D \mathbf{y}, \quad \text{where} \quad D = \begin{pmatrix} a & 0 \\ 0 & \frac{ac - b^2}{a} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \quad (3.72)$$

Anticipating the final result, the equations (3.69) connecting  $\mathbf{x}$  and  $\mathbf{y}$  can themselves be written in matrix form as

$$\mathbf{y} = L^T \mathbf{x} \quad \text{or} \quad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + \frac{b}{a} x_2 \\ x_2 \end{pmatrix}, \quad \text{where} \quad L^T = \begin{pmatrix} 1 & \frac{b}{a} \\ 0 & 1 \end{pmatrix}.$$

Substituting into (3.72), we obtain

$$\mathbf{y}^T D \mathbf{y} = (L^T \mathbf{x})^T D (L^T \mathbf{x}) = \mathbf{x}^T L D L^T \mathbf{x} = \mathbf{x}^T K \mathbf{x}, \quad \text{where} \quad K = L D L^T. \quad (3.73)$$

The result is the *same* factorization (1.61) of the coefficient matrix that we previously obtained via Gaussian Elimination. We are thus led to the realization that *completing the square is the same as the  $L D L^T$  factorization of a symmetric matrix!*

Recall the definition of a regular matrix as one that can be reduced to upper triangular form without any row interchanges. Theorem 1.34 says that the regular symmetric matrices are precisely those that admit an  $L D L^T$  factorization. The identity (3.73) is therefore valid for all regular  $n \times n$  symmetric matrices, and shows how to write the associated quadratic form as a sum of squares:

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \mathbf{y}^T D \mathbf{y} = d_1 y_1^2 + \cdots + d_n y_n^2, \quad \text{where} \quad \mathbf{y} = L^T \mathbf{x}. \quad (3.74)$$

The coefficients  $d_i$  are the diagonal entries of  $D$ , which are the pivots of  $K$ . Furthermore, the diagonal quadratic form is positive definite,  $\mathbf{y}^T D \mathbf{y} > 0$  for all  $\mathbf{y} \neq \mathbf{0}$ , if and only if all the pivots are positive,  $d_i > 0$ . Invertibility of  $L^T$  tells us that  $\mathbf{y} = \mathbf{0}$  if and only if  $\mathbf{x} = \mathbf{0}$ , and hence, positivity of the pivots is equivalent to positive definiteness of the original quadratic form:  $q(\mathbf{x}) > 0$  for all  $\mathbf{x} \neq \mathbf{0}$ . We have thus almost proved the main result that completely characterizes positive definite matrices.

**Theorem 3.43.** A symmetric matrix is positive definite if and only if it is regular and has all positive pivots.

Equivalently, a square matrix  $K$  is positive definite if and only if it can be factored  $K = LDL^T$ , where  $L$  is lower unitriangular and  $D$  is diagonal with all positive diagonal entries.

**Example 3.44.** Consider the symmetric matrix  $K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix}$ . Gaussian Elimination produces the factors

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}, \quad L^T = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

in its factorization  $K = LDL^T$ . Since the pivots — the diagonal entries 1, 2, and 6 in  $D$  — are all positive, Theorem 3.43 implies that  $K$  is positive definite, which means that the associated quadratic form satisfies

$$q(\mathbf{x}) = x_1^2 + 4x_1x_2 - 2x_1x_3 + 6x_2^2 + 9x_3^2 > 0, \quad \text{for all } \mathbf{x} = (x_1, x_2, x_3)^T \neq \mathbf{0}.$$

Indeed, the  $LDL^T$  factorization implies that  $q(\mathbf{x})$  can be explicitly written as a sum of squares:

$$q(\mathbf{x}) = x_1^2 + 4x_1x_2 - 2x_1x_3 + 6x_2^2 + 9x_3^2 = y_1^2 + 2y_2^2 + 6y_3^2, \quad (3.75)$$

where

$$y_1 = x_1 + 2x_2 - x_3, \quad y_2 = x_2 + x_3, \quad y_3 = x_3,$$

are the entries of  $\mathbf{y} = L^T\mathbf{x}$ . Positivity of the coefficients of the  $y_i^2$  (which are the pivots) implies that  $q(\mathbf{x})$  is positive definite.

**Example 3.45.** Let's test whether the matrix  $K = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 7 \\ 3 & 7 & 8 \end{pmatrix}$  is positive definite.

When we perform Gaussian Elimination, the second pivot turns out to be  $-1$ , which immediately implies that  $K$  is not positive definite — even though all its entries are positive. (The third pivot is 3, but this does not affect the conclusion; all it takes is one non-positive pivot to disqualify a matrix from being positive definite. Also, row interchanges aren't of any help, since we are not allowed to perform them when checking for positive definiteness.) This means that the associated quadratic form

$$q(\mathbf{x}) = x_1^2 + 4x_1x_2 + 6x_1x_3 + 3x_2^2 + 14x_2x_3 + 8x_3^2$$

assumes negative values at some points. For instance,  $q(-2, 1, 0) = -1$ .

A direct method for completing the square in a quadratic form goes as follows: The first step is to put all the terms involving  $x_1$  in a suitable square, at the expense of introducing extra terms involving only the other variables. For instance, in the case of the quadratic form in (3.75), the terms involving  $x_1$  can be written as

$$x_1^2 + 4x_1x_2 - 2x_1x_3 = (x_1 + 2x_2 - x_3)^2 - 4x_2^2 + 4x_2x_3 - x_3^2.$$

Therefore,

$$q(\mathbf{x}) = (x_1 + 2x_2 - x_3)^2 + 2x_2^2 + 4x_2x_3 + 8x_3^2 = (x_1 + 2x_2 - x_3)^2 + \tilde{q}(x_2, x_3),$$

where

$$\tilde{q}(x_2, x_3) = 2x_2^2 + 4x_2x_3 + 8x_3^2$$

is a quadratic form that involves only  $x_2, x_3$ . We then repeat the process, combining all the terms involving  $x_2$  in the remaining quadratic form into a square, writing

$$\tilde{q}(x_2, x_3) = 2(x_2 + x_3)^2 + 6x_3^2.$$

This gives the final form

$$q(\mathbf{x}) = (x_1 + 2x_2 - x_3)^2 + 2(x_2 + x_3)^2 + 6x_3^2,$$

which reproduces (3.75).

In general, as long as  $k_{11} \neq 0$ , we can write

$$\begin{aligned} q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} &= k_{11} x_1^2 + 2k_{12} x_1 x_2 + \cdots + 2k_{1n} x_1 x_n + k_{22} x_2^2 + \cdots + k_{nn} x_n^2 \\ &= k_{11} \left( x_1 + \frac{k_{12}}{k_{11}} x_2 + \cdots + \frac{k_{1n}}{k_{11}} x_n \right)^2 + \tilde{q}(x_2, \dots, x_n) \\ &= k_{11} (x_1 + l_{21} x_2 + \cdots + l_{n1} x_n)^2 + \tilde{q}(x_2, \dots, x_n), \end{aligned} \quad (3.76)$$

where

$$l_{21} = \frac{k_{21}}{k_{11}} = \frac{k_{12}}{k_{11}}, \quad \dots \quad l_{n1} = \frac{k_{n1}}{k_{11}} = \frac{k_{1n}}{k_{11}}$$

are precisely the multiples appearing in the matrix  $L$  obtained from applying Gaussian Elimination to  $K$ , while

$$\tilde{q}(x_2, \dots, x_n) = \sum_{i,j=2}^n \tilde{k}_{ij} x_i x_j$$

is a quadratic form involving one less variable. The entries of its symmetric coefficient matrix  $\tilde{K}$  are

$$\tilde{k}_{ij} = \tilde{k}_{ji} = k_{ij} - l_{j1} k_{1i} = k_{ij} - \frac{k_{1j} k_{1i}}{k_{11}}, \quad i, j = 2, \dots, n,$$

which are exactly the same as the entries appearing below and to the right of the first pivot after applying the first phase of the Gaussian Elimination process to  $K$ . In particular, the second pivot of  $K$  is the diagonal entry  $\tilde{k}_{22}$ . We continue by applying the same procedure to the reduced quadratic form  $\tilde{q}(x_2, \dots, x_n)$  and repeating until only the final variable remains. Completing the square at each stage reproduces the corresponding phase of the Gaussian Elimination process. The final result is our formula (3.74) rewriting the original quadratic form as a sum of squares whose coefficients are the pivots.

With this in hand, we can now complete the proof of Theorem 3.43. First, if the upper left entry  $k_{11}$ , namely the first pivot, is not strictly positive, then  $K$  cannot be positive definite, because  $q(\mathbf{e}_1) = \mathbf{e}_1^T K \mathbf{e}_1 = k_{11} \leq 0$ . Otherwise, suppose  $k_{11} > 0$ , and so we can write  $q(\mathbf{x})$  in the form (3.76). We claim that  $q(\mathbf{x})$  is positive definite if and only if the reduced quadratic form  $\tilde{q}(x_2, \dots, x_n)$  is positive definite. Indeed, if  $\tilde{q}$  is positive definite and  $k_{11} > 0$ , then  $q(\mathbf{x})$  is the sum of two positive quantities, which simultaneously vanish if and only if  $x_1 = x_2 = \cdots = x_n = 0$ . On the other hand, suppose  $\tilde{q}(x_2^*, \dots, x_n^*) \leq 0$  for some  $x_2^*, \dots, x_n^*$ , not all zero. Setting  $x_1^* = -l_{21} x_2^* - \cdots - l_{n1} x_n^*$  makes the initial square term in (3.76) equal to 0, so

$$q(x_1^*, x_2^*, \dots, x_n^*) = \tilde{q}(x_2^*, \dots, x_n^*) \leq 0,$$

proving the claim. In particular, positive definiteness of  $\tilde{q}$  requires that the second pivot satisfy  $\tilde{k}_{22} > 0$ . We then continue the reduction procedure outlined in the preceding

paragraph; if a non-positive entry appears in the diagonal pivot position at any stage, the original quadratic form and matrix cannot be positive definite. On the other hand, finding all positive pivots (without using any row interchanges) will, in the absence of numerical errors, ensure positive definiteness.

*Q.E.D.*

## Exercises

- 3.5.1. Are the following matrices are positive definite? (a)  $\begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -2 \\ 1 & -2 & 4 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}$ , (f)  $\begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}$ .

- 3.5.2. Find an  $LDL^T$  factorization of the following symmetric matrices. Which are positive

- definite? (a)  $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$ , (b)  $\begin{pmatrix} 5 & -1 \\ -1 & 3 \end{pmatrix}$ , (c)  $\begin{pmatrix} 3 & -1 & 3 \\ -1 & 5 & 1 \\ 3 & 1 & 5 \end{pmatrix}$ , (d)  $\begin{pmatrix} -2 & 1 & -1 \\ 1 & -2 & 1 \\ -1 & 1 & -2 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & 1 & -2 \\ 1 & 1 & -3 \\ -2 & -3 & 11 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}$ , (g)  $\begin{pmatrix} 3 & 2 & 1 & 0 \\ 2 & 3 & 0 & 1 \\ 1 & 0 & 3 & 2 \\ 0 & 1 & 2 & 3 \end{pmatrix}$ , (h)  $\begin{pmatrix} 2 & 1 & -2 & 0 \\ 1 & 1 & -3 & 2 \\ -2 & -3 & 10 & -1 \\ 0 & 2 & -1 & 7 \end{pmatrix}$ .

- 3.5.3. (a) For which values of  $c$  is the matrix  $A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & c & 1 \\ 0 & 1 & 1 \end{pmatrix}$  positive definite? (b) For the

particular value  $c = 3$ , carry out elimination to find the factorization  $A = LDL^T$ . (c) Use your result from part (b) to rewrite the quadratic form  $q(x, y, z) = x^2 + 2xy + 3y^2 + 2yz + z^2$  as a sum of squares. (d) Explain how your result is related to the positive definiteness of  $A$ .

- 3.5.4. Write the quadratic form  $q(\mathbf{x}) = x_1^2 + x_1 x_2 + 2x_2^2 - x_1 x_3 + 3x_3^2$  in the form  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$  for some symmetric matrix  $K$ . Is  $q(\mathbf{x})$  positive definite?

- 3.5.5. Write the following quadratic forms on  $\mathbb{R}^2$  as a sum of squares. Which are positive definite? (a)  $x^2 + 8xy + y^2$ , (b)  $x^2 - 4xy + 7y^2$ , (c)  $x^2 - 2xy - y^2$ , (d)  $x^2 + 6xy$ .

- 3.5.6. Prove that the following quadratic forms on  $\mathbb{R}^3$  are positive definite by writing each as a sum of squares: (a)  $x^2 + 4xz + 3y^2 + 5z^2$ , (b)  $x^2 + 3xy + 3y^2 - 2xz + 8z^2$ , (c)  $2x_1^2 + x_1 x_2 - 2x_1 x_3 + 2x_2^2 - 2x_2 x_3 + 2x_3^2$ .

- 3.5.7. Write the following quadratic forms in matrix notation and determine if they are positive definite: (a)  $x^2 + 4xz + 2y^2 + 8yz + 12z^2$ , (b)  $3x^2 - 2y^2 - 8xy + xz + z^2$ , (c)  $x^2 + 2xy + 2y^2 - 4xz - 6yz + 6z^2$ , (d)  $3x_1^2 - x_2^2 + 5x_3^2 + 4x_1 x_2 - 7x_1 x_3 + 9x_2 x_3$ , (e)  $x_1^2 + 4x_1 x_2 - 2x_1 x_3 + 5x_2^2 - 2x_2 x_4 + 6x_3^2 - x_3 x_4 + 4x_4^2$ .

- 3.5.8. For what values of  $a, b$ , and  $c$  is the quadratic form  $x^2 + axy + y^2 + bxz + cyz + z^2$  positive definite?

- 3.5.9. *True or false:* Every planar quadratic form  $q(x, y) = ax^2 + 2bxy + cy^2$  can be written as a sum of squares.

- 3.5.10. (a) Prove that a positive definite matrix has positive determinant:  $\det K > 0$ . (b) Show that a positive definite matrix has positive trace:  $\text{tr } K > 0$ . (c) Show that every  $2 \times 2$  symmetric matrix with positive determinant and positive trace is positive definite. (d) Find a symmetric  $3 \times 3$  matrix with positive determinant and positive trace that is not positive definite.

3.5.11. (a) Prove that if  $K_1, K_2$  are positive definite  $n \times n$  matrices, then  $K = \begin{pmatrix} K_1 & \mathbf{0} \\ \mathbf{0} & K_2 \end{pmatrix}$  is a positive definite  $2n \times 2n$  matrix. (b) Is the converse true?

3.5.12. Let  $\|\cdot\|$  be any norm on  $\mathbb{R}^n$ . (a) Show that  $q(\mathbf{x})$  is a positive definite quadratic form if and only if  $q(\mathbf{u}) > 0$  for all unit vectors,  $\|\mathbf{u}\| = 1$ . (b) Prove that if  $S = S^T$  is any symmetric matrix, then  $K = S + c \mathbf{I} > 0$  is positive definite if  $c$  is sufficiently large.

3.5.13. Prove that every symmetric matrix  $S = K + N$  can be written as the sum of a positive definite matrix  $K$  and a negative definite matrix  $N$ . *Hint:* Use Exercise 3.5.12(b).

◇ 3.5.14. (a) Prove that every regular symmetric matrix can be decomposed as a linear combination

$$K = d_1 \mathbf{l}_1 \mathbf{l}_1^T + d_2 \mathbf{l}_2 \mathbf{l}_2^T + \cdots + d_n \mathbf{l}_n \mathbf{l}_n^T \quad (3.77)$$

of symmetric rank 1 matrices, as in Exercise 1.8.15, where  $\mathbf{l}_1, \dots, \mathbf{l}_n$  are the columns of the lower unitriangular matrix  $L$  and  $d_1, \dots, d_n$  are the pivots, i.e., the diagonal entries of  $D$ .

*Hint:* See Exercise 1.2.34. (b) Decompose  $\begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}$  and  $\begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$  in this manner.

◇ 3.5.15. There is an alternative criterion for positive definiteness based on subdeterminants of the matrix. The  $2 \times 2$  version already appears in (3.70). (a) Prove that a  $3 \times 3$  matrix

$K = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$  is positive definite if and only if  $a > 0$ ,  $ad - b^2 > 0$ , and  $\det K > 0$ .

(b) Prove the general version: an  $n \times n$  matrix  $K > 0$  is positive definite if and only if its upper left square  $k \times k$  submatrices have positive determinant for all  $k = 1, \dots, n$ .

*Hint:* See Exercise 1.9.17.

◇ 3.5.16. Let  $K$  be a symmetric matrix. Prove that if a non-positive diagonal entry appears anywhere (not necessarily in the pivot position) in the matrix during Regular Gaussian Elimination, then  $K$  is not positive definite.

◇ 3.5.17. Formulate a determinantal criterion similar to that in Exercise 3.5.15 for negative definite matrices. Write out the  $2 \times 2$  and  $3 \times 3$  cases explicitly.

3.5.18. *True or false:* A negative definite matrix must have negative trace and negative determinant.

## The Cholesky Factorization

The identity (3.73) shows us how to write an arbitrary regular quadratic form  $q(\mathbf{x})$  as a linear combination of squares. We can push this result slightly further in the positive definite case. Since each pivot  $d_i$  is positive, we can write the diagonal quadratic form (3.74) as a sum of pure squares:

$$d_1 y_1^2 + \cdots + d_n y_n^2 = (\sqrt{d_1} y_1)^2 + \cdots + (\sqrt{d_n} y_n)^2 = z_1^2 + \cdots + z_n^2,$$

where  $z_i = \sqrt{d_i} y_i$ . In matrix form, we are writing

$$\hat{q}(\mathbf{y}) = \mathbf{y}^T D \mathbf{y} = \mathbf{z}^T \mathbf{z} = \|\mathbf{z}\|^2, \quad \text{where } \mathbf{z} = S \mathbf{y}, \quad \text{with } S = \text{diag} \left( \sqrt{d_1}, \dots, \sqrt{d_n} \right).$$

Since  $D = S^2$ , the matrix  $S$  can be thought of as a “square root” of the diagonal matrix  $D$ . Substituting back into (1.58), we deduce the *Cholesky factorization*

$$K = LDL^T = LSS^T L^T = MM^T, \quad \text{where } M = LS, \quad (3.78)$$

of a positive definite matrix, first proposed by the early twentieth-century French geographer André-Louis Cholesky for solving problems in geodetic surveying. Note that  $M$  is a

lower triangular matrix with all positive diagonal entries, namely the square roots of the pivots:  $m_{ii} = \sqrt{d_i}$ . Applying the Cholesky factorization to the corresponding quadratic form produces

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \mathbf{x}^T M M^T \mathbf{x} = \mathbf{z}^T \mathbf{z} = \|\mathbf{z}\|^2, \quad \text{where} \quad \mathbf{z} = M^T \mathbf{x}. \quad (3.79)$$

We can interpret (3.79) as a change of variables from  $\mathbf{x}$  to  $\mathbf{z}$  that converts an arbitrary inner product norm, as defined by the square root of the positive definite quadratic form  $q(\mathbf{x})$ , into the standard Euclidean norm  $\|\mathbf{z}\|$ .

**Example 3.46.** For the matrix  $K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix}$  considered in Example 3.44, the

Cholesky formula (3.78) gives  $K = M M^T$ , where

$$M = L S = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{6} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & \sqrt{2} & 0 \\ -1 & \sqrt{2} & \sqrt{6} \end{pmatrix}.$$

The associated quadratic function can then be written as a sum of pure squares:

$$q(\mathbf{x}) = x_1^2 + 4x_1 x_2 - 2x_1 x_3 + 6x_2^2 + 9x_3^2 = z_1^2 + z_2^2 + z_3^2,$$

where

$$\mathbf{z} = M^T \mathbf{x}, \quad \text{or, explicitly,} \quad z_1 = x_1 + 2x_2 - x_3, \quad z_2 = \sqrt{2}x_2 + \sqrt{2}x_3, \quad z_3 = \sqrt{6}x_3.$$

## Exercises

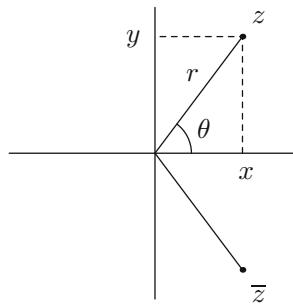
- 3.5.19. Find the Cholesky factorizations of the following matrices: (a)  $\begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}$ ,  
 (b)  $\begin{pmatrix} 4 & -12 \\ -12 & 45 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -2 \\ 1 & -2 & 14 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$ .

- 3.5.20. Which of the matrices in Exercise 3.5.1 have a Cholesky factorization? For those that do, write out the factorization.

- 3.5.21. Write the following positive definite quadratic forms as a sum of pure squares, as in (3.79): (a)  $16x_1^2 + 25x_2^2$ , (b)  $x_1^2 - 2x_1 x_2 + 4x_2^2$ , (c)  $5x_1^2 + 4x_1 x_2 + 3x_2^2$ ,  
 (d)  $3x_1^2 - 2x_1 x_2 - 2x_1 x_3 + 2x_2^2 + 6x_3^2$ , (e)  $x_1^2 + x_1 x_2 + x_2^2 + x_2 x_3 + x_3^2$ ,  
 (f)  $4x_1^2 - 2x_1 x_2 - 4x_1 x_3 + \frac{1}{2}x_2^2 - x_2 x_3 + 6x_3^2$ ,  
 (g)  $3x_1^2 + 2x_1 x_2 + 3x_2^2 + 2x_2 x_3 + 3x_3^2 + 2x_3 x_4 + 3x_4^2$ .

## 3.6 Complex Vector Spaces

Although physical applications ultimately require real answers, complex numbers and complex vector spaces play an extremely useful, if not essential, role in the intervening analysis. Particularly in the description of periodic phenomena, complex numbers and complex exponentials help to simplify complicated trigonometric formulas. Complex variable methods



**Figure 3.7.** Complex Numbers.

are ubiquitous in electrical engineering, Fourier analysis, potential theory, fluid mechanics, electromagnetism, and many other applied fields, [49, 79]. In quantum mechanics, the basic physical quantities are complex-valued wave functions, [54]. Moreover, the Schrödinger equation, which governs quantum dynamics, is an inherently complex partial differential equation.

In this section, we survey the principal properties of complex numbers and complex vector spaces. Most of the constructions are straightforward adaptations of their real counterparts, and so will not be dwelled on at length. The one exception is the complex version of an inner product, which does introduce some novelties not found in its simpler real sibling.

### Complex Numbers

Recall that a *complex number* is an expression of the form  $z = x + iy$ , where  $x, y \in \mathbb{R}$  are real and<sup>†</sup>  $i = \sqrt{-1}$ . The set of all complex numbers (scalars) is denoted by  $\mathbb{C}$ . We call  $x = \operatorname{Re} z$  the *real part* of  $z$  and  $y = \operatorname{Im} z$  the *imaginary part* of  $z = x + iy$ . (Note: The imaginary part is the real number  $y$ , *not*  $iy$ .) A real number  $x$  is merely a complex number with zero imaginary part,  $\operatorname{Im} z = 0$ , and so we may regard  $\mathbb{R} \subset \mathbb{C}$ . Complex addition and multiplication are based on simple adaptations of the rules of real arithmetic to include the identity  $i^2 = -1$ , and so

$$\begin{aligned} (x + iy) + (u + iv) &= (x + u) + i(y + v), \\ (x + iy)(u + iv) &= (xu - yv) + i(xv + yu). \end{aligned} \tag{3.80}$$

Complex numbers enjoy all the usual laws of real addition and multiplication, *including commutativity*:  $zw = wz$ .

We can identify a complex number  $x + iy$  with a vector  $(x, y)^T \in \mathbb{R}^2$  in the real plane. For this reason,  $\mathbb{C}$  is sometimes referred to as the *complex plane*. Complex addition (3.80) corresponds to vector addition, but complex multiplication does not have a readily identifiable vector counterpart.

Another useful operation on complex numbers is that of complex conjugation.

**Definition 3.47.** The *complex conjugate* of  $z = x + iy$  is  $\bar{z} = x - iy$ , whereby  $\operatorname{Re} \bar{z} = \operatorname{Re} z$ , while  $\operatorname{Im} \bar{z} = -\operatorname{Im} z$ .

<sup>†</sup> To avoid confusion with the symbol for current, electrical engineers prefer to use  $j$  to indicate the imaginary unit.

Geometrically, the complex conjugate of  $z$  is obtained by reflecting the corresponding vector through the real axis, as illustrated in [Figure 3.7](#). In particular  $\bar{z} = z$  if and only if  $z$  is real. Note that

$$\operatorname{Re} z = \frac{z + \bar{z}}{2}, \quad \operatorname{Im} z = \frac{z - \bar{z}}{2i}. \quad (3.81)$$

Complex conjugation is compatible with complex arithmetic:

$$\overline{z + w} = \bar{z} + \bar{w}, \quad \overline{zw} = \bar{z} \bar{w}.$$

In particular, the product of a complex number and its conjugate,

$$z\bar{z} = (x + iy)(x - iy) = x^2 + y^2, \quad (3.82)$$

is real and non-negative. Its square root is known as the *modulus* or *norm* of the complex number  $z = x + iy$ , and written

$$|z| = \sqrt{x^2 + y^2}. \quad (3.83)$$

Note that  $|z| \geq 0$ , with  $|z| = 0$  if and only if  $z = 0$ . The modulus  $|z|$  generalizes the absolute value of a real number, and coincides with the standard Euclidean norm in the  $xy$ -plane, which implies the validity of the triangle inequality

$$|z + w| \leq |z| + |w|. \quad (3.84)$$

Equation (3.82) can be rewritten in terms of the modulus as

$$z\bar{z} = |z|^2. \quad (3.85)$$

Rearranging the factors, we deduce the formula for the reciprocal of a nonzero complex number:

$$\frac{1}{z} = \frac{\bar{z}}{|z|^2}, \quad z \neq 0, \quad \text{or, equivalently,} \quad \frac{1}{x + iy} = \frac{x - iy}{x^2 + y^2}. \quad (3.86)$$

The general formula for complex division,

$$\frac{w}{z} = \frac{w\bar{z}}{|z|^2} \quad \text{or} \quad \frac{u + iv}{x + iy} = \frac{(xu + yv) + i(xv - yu)}{x^2 + y^2}, \quad (3.87)$$

is an immediate consequence.

The modulus of a complex number,

$$r = |z| = \sqrt{x^2 + y^2},$$

is one component of its polar coordinate representation

$$x = r \cos \theta, \quad y = r \sin \theta \quad \text{or} \quad z = r(\cos \theta + i \sin \theta). \quad (3.88)$$

The polar angle, which measures the angle that the line connecting  $z$  to the origin makes with the horizontal axis, is known as the *phase*, and written

$$\theta = \operatorname{ph} z. \quad (3.89)$$

As such, the phase is defined only up to an integer multiple of  $2\pi$ . The more common term for the angle is the *argument*, written  $\arg z = \operatorname{ph} z$ . However, we prefer to use “phase” throughout this text, in part to avoid confusion with the argument  $z$  of a function  $f(z)$ .

We note that the modulus and phase of a product of complex numbers can be readily computed:

$$|zw| = |z||w|, \quad \text{ph}(zw) = \text{ph } z + \text{ph } w. \quad (3.90)$$

Complex conjugation preserves the modulus, but reverses the sign of the phase:

$$|\bar{z}| = |z|, \quad \text{ph } \bar{z} = -\text{ph } z. \quad (3.91)$$

One of the most profound formulas in all of mathematics is *Euler's formula*

$$e^{i\theta} = \cos \theta + i \sin \theta, \quad (3.92)$$

relating the complex exponential with the real sine and cosine functions. It has a variety of mathematical justifications; see Exercise 3.6.23 for one that is based on comparing power series. Euler's formula can be used to compactly rewrite the polar form (3.88) of a complex number as

$$z = r e^{i\theta} \quad \text{where} \quad r = |z|, \quad \theta = \text{ph } z. \quad (3.93)$$

The complex conjugation identity

$$e^{-i\theta} = \cos(-\theta) + i \sin(-\theta) = \cos \theta - i \sin \theta = \overline{e^{i\theta}}$$

permits us to express the basic trigonometric functions in terms of complex exponentials:

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}. \quad (3.94)$$

These formulas are very useful when working with trigonometric identities and integrals.

The exponential of a general complex number is easily derived from the Euler formula and the standard properties of the exponential function — which carry over unaltered to the complex domain; thus,

$$e^z = e^{x+i y} = e^x e^{i y} = e^x \cos y + i e^x \sin y. \quad (3.95)$$

Note that  $e^{2\pi i} = 1$ , and hence the exponential function is periodic,

$$e^{z+2\pi i} = e^z, \quad (3.96)$$

with imaginary period  $2\pi i$  — indicative of the periodicity of the trigonometric functions in Euler's formula.

## Exercises

3.6.1. Write down a single equation that relates the five most important numbers in mathematics, which are 0, 1,  $e$ ,  $\pi$ , and  $i$ .

3.6.2. For any integer  $k$ , prove that  $e^{k\pi i} = (-1)^k$ .

3.6.3. Is the formula  $1^z = 1$  valid for all complex values of  $z$ ?

3.6.4. What is wrong with the calculation  $e^{2a\pi i} = (e^{2\pi i})^a = 1^a = 1$ ?

3.6.5. (a) Write  $i$  in phase-modulus form. (b) Use this expression to find  $\sqrt{i}$ , i.e., a complex number  $z$  such that  $z^2 = i$ . Can you find a second square root? (c) Find explicit formulas for the three third roots and four fourth roots of  $i$ .

3.6.6. In Figure 3.7, where would you place the point  $1/z$ ?

3.6.7. (a) If  $z$  moves counterclockwise around a circle of radius  $r$  in the complex plane, around which circle and in which direction does  $w = 1/z$  move? (b) What about  $w = \bar{z}$ ?  
 (c) What if the circle is not centered at the origin?

◇ 3.6.8. Show that  $-|z| \leq \operatorname{Re} z \leq |z|$  and  $-|z| \leq \operatorname{Im} z \leq |z|$ .

◇ 3.6.9. Prove that if  $\varphi$  is real, then  $\operatorname{Re}(e^{i\varphi} z) \leq |z|$ , with equality if and only if  $\varphi = -\operatorname{ph} z$ .

3.6.10. Prove the identities in (3.90) and (3.91).

3.6.11. Prove  $\operatorname{ph}(z/w) = \operatorname{ph} z - \operatorname{ph} w = \operatorname{ph}(z\bar{w})$  is equal to the angle between the vectors representing  $z$  and  $w$ .

3.6.12. The phase of a complex number  $z = x + iy$  is often written as  $\operatorname{ph} z = \tan^{-1}(y/x)$ . Explain why this formula is ambiguous, and does not uniquely define  $\operatorname{ph} z$ .

3.6.13. Show that if we identify the complex numbers  $z, w$  with vectors in the plane, then their Euclidean dot product is equal to  $\operatorname{Re}(z\bar{w})$ .

3.6.14. (a) Prove that the complex numbers  $z$  and  $w$  correspond to orthogonal vectors in  $\mathbb{R}^2$  if and only if  $\operatorname{Re} z\bar{w} = 0$ . (b) Prove that  $z$  and  $iz$  are always orthogonal.

3.6.15. Prove that  $e^{z+w} = e^z e^w$ . Conclude that  $e^{mz} = (e^z)^m$  whenever  $m$  is an integer.

3.6.16. (a) Use the formula  $e^{2i\theta} = (e^{i\theta})^2$  to deduce the well-known trigonometric identities for  $\cos 2\theta$  and  $\sin 2\theta$ . (b) Derive the corresponding identities for  $\cos 3\theta$  and  $\sin 3\theta$ .  
 (c) Write down the explicit identities for  $\cos m\theta$  and  $\sin m\theta$  as polynomials in  $\cos \theta$  and  $\sin \theta$ . Hint: Apply the Binomial Formula to  $(e^{i\theta})^m$ .

◇ 3.6.17. Use complex exponentials to prove the identity  $\cos \theta - \cos \varphi = 2 \cos \frac{\theta - \varphi}{2} \cos \frac{\theta + \varphi}{2}$ .

3.6.18. Prove that if  $z = x + iy$ , then  $|e^z| = e^x$ ,  $\operatorname{ph} e^z = y$ .

3.6.19. The formulas  $\cos z = \frac{e^{iz} + e^{-iz}}{2}$  and  $\sin z = \frac{e^{iz} - e^{-iz}}{2i}$  serve to define the basic

complex trigonometric functions. Write out the formulas for their real and imaginary parts in terms of  $z = x + iy$ , and show that  $\cos z$  and  $\sin z$  reduce to their usual real forms when  $z = x$  is real. What do they become when  $z = iy$  is purely imaginary?

3.6.20. The complex *hyperbolic functions* are defined as  $\cosh z = \frac{e^z + e^{-z}}{2}$ ,  $\sinh z = \frac{e^z - e^{-z}}{2}$ .

(a) Write out the formulas for their real and imaginary parts in terms of  $z = x + iy$ .

(b) Prove that  $\cos iz = \cosh z$  and  $\sin iz = i \sinh z$ .

◇ 3.6.21. Generalizing Example 2.17c, by a *trigonometric polynomial* of degree  $\leq n$ , we mean a function  $T(x) = \sum_{0 \leq j+k \leq n} c_{jk} (\cos \theta)^j (\sin \theta)^k$  in the powers of the sine and cosine functions up to degree  $n$ . (a) Use formula (3.94) to prove that every trigonometric polynomial of degree  $\leq n$  can be written as a complex linear combination of the  $2n+1$  complex exponentials  $e^{-n i \theta}, \dots, e^{-i \theta}, e^{0 i \theta} = 1, e^{i \theta}, e^{2 i \theta}, \dots, e^{n i \theta}$ . (b) Prove that every trigonometric polynomial of degree  $\leq n$  can be written as a real linear combination of the trigonometric functions  $1, \cos \theta, \sin \theta, \cos 2\theta, \sin 2\theta, \dots, \cos n\theta, \sin n\theta$ .  
 (c) Write out the following trigonometric polynomials in both of the preceding forms:

(i)  $\cos^2 \theta$ , (ii)  $\cos \theta \sin \theta$ , (iii)  $\cos^3 \theta$ , (iv)  $\sin^4 \theta$ , (v)  $\cos^2 \theta \sin^2 \theta$ .

◇ 3.6.22. Write out the real and imaginary parts of the power function  $x^c$  with complex exponent  $c = a + ib \in \mathbb{C}$ .

◇ 3.6.23. Write the power series expansions for  $e^{ix}$ . Prove that the real terms give the power series for  $\cos x$ , while the imaginary terms give that of  $\sin x$ . Use this identification to justify Euler's formula (3.92).

- ◇ 3.6.24. The derivative of a complex-valued function  $f(x) = u(x) + i v(x)$ , depending on a real variable  $x$ , is given by  $f'(x) = u'(x) + i v'(x)$ . (a) Prove that if  $\lambda = \mu + i \nu$  is any complex scalar, then  $\frac{d}{dx} e^{\lambda x} = \lambda e^{\lambda x}$ . (b) Prove, conversely,  $\int_a^b e^{\lambda x} dx = \frac{1}{\lambda} (e^{\lambda b} - e^{\lambda a})$  provided  $\lambda \neq 0$ .
- 3.6.25. Use the complex trigonometric formulas (3.94) and Exercise 3.6.24 to evaluate the following trigonometric integrals: (a)  $\int \cos^2 x dx$ , (b)  $\int \sin^2 x dx$ , (c)  $\int \cos x \sin x dx$ , (d)  $\int \cos 3x \sin 5x dx$ . How did you calculate them in first-year calculus? If you're not convinced this method is easier, try the more complicated integrals (e)  $\int \cos^4 x dx$ , (f)  $\int \sin^4 x dx$ , (g)  $\int \cos^2 x \sin^2 x dx$ , (h)  $\int \cos 3x \sin 5x \cos 7x dx$ .

## Complex Vector Spaces and Inner Products

A *complex vector space* is defined in exactly the same manner as its real counterpart, as in Definition 2.1, the only difference being that we replace real scalars by complex scalars. The most basic example is the  $n$ -dimensional complex vector space  $\mathbb{C}^n$  consisting of all column vectors  $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$  that have  $n$  complex entries  $z_1, \dots, z_n \in \mathbb{C}$ . Vector addition and scalar multiplication are defined in the obvious manner, and verification of each of the vector space axioms is immediate.

We can write any complex vector  $\mathbf{z} = \mathbf{x} + i \mathbf{y} \in \mathbb{C}^n$  as a linear combination of two real vectors  $\mathbf{x} = \operatorname{Re} \mathbf{z}$  and  $\mathbf{y} = \operatorname{Im} \mathbf{z} \in \mathbb{R}^n$  called its *real* and *imaginary parts*. Its complex conjugate  $\bar{\mathbf{z}} = \mathbf{x} - i \mathbf{y}$  is obtained by taking the complex conjugates of its individual entries. Thus, for example, if

$$\mathbf{z} = \begin{pmatrix} 1 + 2i \\ -3 \\ 5i \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix} + i \begin{pmatrix} 2 \\ 0 \\ 5 \end{pmatrix}, \quad \text{then} \quad \operatorname{Re} \mathbf{z} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}, \quad \operatorname{Im} \mathbf{z} = \begin{pmatrix} 2 \\ 0 \\ 5 \end{pmatrix},$$

and so its complex conjugate is  $\bar{\mathbf{z}} = \begin{pmatrix} 1 - 2i \\ -3 \\ -5i \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix} - i \begin{pmatrix} 2 \\ 0 \\ 5 \end{pmatrix}$ .

In particular,  $\mathbf{z} \in \mathbb{R}^n \subset \mathbb{C}^n$  is a real vector if and only if  $\mathbf{z} = \bar{\mathbf{z}}$ .

Most of the vector space concepts we developed in the real domain, including span, linear independence, basis, and dimension, can be straightforwardly extended to the complex regime. The one exception is the concept of an inner product, which requires a little thought. In analysis, the primary applications of inner products and norms rely on the associated inequalities: Cauchy–Schwarz and triangle. But there is no natural ordering of the complex numbers, and so one *cannot* assign a meaning to a complex inequality like  $z < w$ . Inequalities make sense only in the real domain, and so the norm of a complex vector should still be a positive and real. With this in mind, the naïve idea of simply summing the squares of the entries of a complex vector will *not* define a norm on  $\mathbb{C}^n$ , since the result will typically be complex. Moreover, some nonzero complex vectors, e.g.,  $(1, i)^T$ , would then have zero “norm”.

The correct definition is modeled on the formula

$$|z| = \sqrt{z\bar{z}},$$

which defines the modulus of a complex scalar  $z \in \mathbb{C}$ . If, in analogy with the real definition (3.7), the quantity inside the square root should represent the inner product of  $z$  with

itself, then we should define the “dot product” between two complex numbers to be

$$z \cdot w = z\bar{w}, \quad \text{so that} \quad z \cdot z = z\bar{z} = |z|^2.$$

Writing out the formula when  $z = x + iy$  and  $w = u + iv$ , we obtain

$$z \cdot w = z\bar{w} = (x + iy)(u - iv) = (xu + yv) + i(yu - xv). \quad (3.97)$$

Thus, the dot product of two complex numbers is, in general, complex. The real part of  $z \cdot w$  is, in fact, the Euclidean dot product between the corresponding vectors in  $\mathbb{R}^2$ , while its imaginary part is, interestingly, their scalar cross product, cf. (3.22).

The vector version of this construction is named after the nineteenth-century French mathematician Charles Hermite, and called the *Hermitian dot product* on  $\mathbb{C}^n$ . It has the explicit formula

$$\mathbf{z} \cdot \mathbf{w} = \mathbf{z}^T \bar{\mathbf{w}} = z_1 \bar{w}_1 + z_2 \bar{w}_2 + \cdots + z_n \bar{w}_n, \quad \text{for } \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}. \quad (3.98)$$

Pay attention to the fact that we must apply complex conjugation to all the entries of the second vector. For example, if

$$\mathbf{z} = \begin{pmatrix} 1+i \\ 3+2i \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1+2i \\ i \end{pmatrix}, \quad \text{then} \quad \mathbf{z} \cdot \mathbf{w} = (1+i)(1-2i) + (3+2i)(-i) = 5-4i.$$

On the other hand,

$$\mathbf{w} \cdot \mathbf{z} = (1+2i)(1-i) + i(3-2i) = 5+4i,$$

and we conclude that the Hermitian dot product is *not* symmetric. Indeed, reversing the order of the vectors conjugates their dot product:

$$\mathbf{w} \cdot \mathbf{z} = \overline{\mathbf{z} \cdot \mathbf{w}}.$$

This is an unexpected complication, but it does have the desired effect that the induced norm, namely

$$0 \leq \|\mathbf{z}\| = \sqrt{\mathbf{z} \cdot \mathbf{z}} = \sqrt{\mathbf{z}^T \bar{\mathbf{z}}} = \sqrt{|z_1|^2 + \cdots + |z_n|^2}, \quad (3.99)$$

is strictly positive for all  $\mathbf{0} \neq \mathbf{z} \in \mathbb{C}^n$ . For example, if

$$\mathbf{z} = \begin{pmatrix} 1+3i \\ -2i \\ -5 \end{pmatrix}, \quad \text{then} \quad \|\mathbf{z}\| = \sqrt{|1+3i|^2 + |-2i|^2 + |-5|^2} = \sqrt{39}.$$

The Hermitian dot product is well behaved under complex vector addition:

$$(\mathbf{z} + \tilde{\mathbf{z}}) \cdot \mathbf{w} = \mathbf{z} \cdot \mathbf{w} + \tilde{\mathbf{z}} \cdot \mathbf{w}, \quad \mathbf{z} \cdot (\mathbf{w} + \tilde{\mathbf{w}}) = \mathbf{z} \cdot \mathbf{w} + \mathbf{z} \cdot \tilde{\mathbf{w}}.$$

However, while complex scalar multiples can be extracted from the first vector without alteration, when they multiply the second vector, they emerge as complex conjugates:

$$(c\mathbf{z}) \cdot \mathbf{w} = c(\mathbf{z} \cdot \mathbf{w}), \quad \mathbf{z} \cdot (c\mathbf{w}) = \bar{c}(\mathbf{z} \cdot \mathbf{w}), \quad c \in \mathbb{C}.$$

Thus, the Hermitian dot product is not bilinear in the strict sense, but satisfies something that, for lack of a better name, is known as *sesquilinearity*.

The general definition of an inner product on a complex vector space is modeled on the preceding properties of the Hermitian dot product.

**Definition 3.48.** An *inner product* on the complex vector space  $V$  is a pairing that takes two vectors  $\mathbf{v}, \mathbf{w} \in V$  and produces a complex number  $\langle \mathbf{v}, \mathbf{w} \rangle \in \mathbb{C}$ , subject to the following requirements, for  $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ , and  $c, d \in \mathbb{C}$ :

(i) *Sesquilinearity:*

$$\begin{aligned}\langle c\mathbf{u} + d\mathbf{v}, \mathbf{w} \rangle &= c\langle \mathbf{u}, \mathbf{w} \rangle + d\langle \mathbf{v}, \mathbf{w} \rangle, \\ \langle \mathbf{u}, c\mathbf{v} + d\mathbf{w} \rangle &= \bar{c}\langle \mathbf{u}, \mathbf{v} \rangle + \bar{d}\langle \mathbf{u}, \mathbf{w} \rangle.\end{aligned}\quad (3.100)$$

(ii) *Conjugate Symmetry:*

$$\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}. \quad (3.101)$$

(iii) *Positivity:*

$$\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle \geq 0, \quad \text{and} \quad \langle \mathbf{v}, \mathbf{v} \rangle = 0 \quad \text{if and only if} \quad \mathbf{v} = \mathbf{0}. \quad (3.102)$$

Thus, when dealing with a complex inner product space, one must pay careful attention to the complex conjugate that appears when the second argument in the inner product is multiplied by a complex scalar, as well as the complex conjugate that appears when the order of the two arguments is reversed. But, once this initial complication has been properly taken into account, the further properties of the inner product carry over directly from the real domain. Exercise 3.6.45 contains the formula for a general inner product on the complex vector space  $\mathbb{C}^n$ .

**Theorem 3.49.** The Cauchy–Schwarz inequality,

$$|\langle \mathbf{v}, \mathbf{w} \rangle| \leq \|\mathbf{v}\| \|\mathbf{w}\|, \quad (3.103)$$

with  $|\cdot|$  now denoting the complex modulus, and the triangle inequality

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\| \quad (3.104)$$

are both valid on an arbitrary complex inner product space.

The proof of (3.103–104) is modeled on the real case, and the details are left to the reader.

**Example 3.50.** The vectors  $\mathbf{v} = (1 + i, 2i, -3)^T$ ,  $\mathbf{w} = (2 - i, 1, 2 + 2i)^T$ , satisfy

$$\begin{aligned}\|\mathbf{v}\| &= \sqrt{2+4+9} = \sqrt{15}, & \|\mathbf{w}\| &= \sqrt{5+1+8} = \sqrt{14}, \\ \mathbf{v} \cdot \mathbf{w} &= (1+i)(2+i) + 2i + (-3)(2-2i) = -5 + 11i.\end{aligned}$$

Thus, the Cauchy–Schwarz inequality reads

$$|\langle \mathbf{v}, \mathbf{w} \rangle| = |-5 + 11i| = \sqrt{146} \leq \sqrt{210} = \sqrt{15} \sqrt{14} = \|\mathbf{v}\| \|\mathbf{w}\|.$$

Similarly, the triangle inequality tells us that

$$\|\mathbf{v} + \mathbf{w}\| = \|(3, 1+2i, -1+2i)^T\| = \sqrt{9+5+5} = \sqrt{19} \leq \sqrt{15} + \sqrt{14} = \|\mathbf{v}\| + \|\mathbf{w}\|.$$

**Example 3.51.** Let  $C^0[-\pi, \pi]$  denote the complex vector space consisting of all complex-valued continuous functions  $f(x) = u(x) + i v(x)$  depending upon the *real* variable

$-\pi \leq x \leq \pi$ . The Hermitian  $L^2$  inner product on  $C^0[-\pi, \pi]$  is defined as

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx, \quad (3.105)$$

i.e., the integral of  $f$  times the complex conjugate of  $g$ , with corresponding norm

$$\|f\| = \sqrt{\int_{-\pi}^{\pi} |f(x)|^2 dx} = \sqrt{\int_{-\pi}^{\pi} [u(x)^2 + v(x)^2] dx}. \quad (3.106)$$

The reader can verify that (3.105) satisfies the Hermitian inner product axioms.

In particular, if  $k, l$  are integers, then the inner product of the complex exponential functions  $e^{ikx}$  and  $e^{ilx}$  is

$$\langle e^{ikx}, e^{ilx} \rangle = \int_{-\pi}^{\pi} e^{ikx} e^{-ilx} dx = \int_{-\pi}^{\pi} e^{i(k-l)x} dx = \begin{cases} 2\pi, & k = l, \\ \frac{e^{i(k-l)x}}{i(k-l)} \Big|_{x=-\pi}^{\pi} = 0, & k \neq l. \end{cases}$$

We conclude that when  $k \neq l$ , the complex exponentials  $e^{ikx}$  and  $e^{ilx}$  are orthogonal, since their inner product is zero. The complex formulation of Fourier analysis, [61, 77], is founded on this key example.

## Exercises

3.6.26. Determine whether the indicated sets of complex vectors are linearly independent or

dependent. (a)  $\begin{pmatrix} i \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ i \end{pmatrix}$ , (b)  $\begin{pmatrix} 1+i \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 2 \\ 1-i \end{pmatrix}$ , (c)  $\begin{pmatrix} 1+3i \\ 2-i \end{pmatrix}$ ,  $\begin{pmatrix} 2-3i \\ 1-i \end{pmatrix}$ ,  
 (d)  $\begin{pmatrix} -2+i \\ i \end{pmatrix}$ ,  $\begin{pmatrix} 4-3i \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 2i \\ 1-5i \end{pmatrix}$ , (e)  $\begin{pmatrix} 1+2i \\ 2 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 2 \\ 0 \\ 1-i \end{pmatrix}$ ,  
 (f)  $\begin{pmatrix} 1 \\ 3i \\ 2-i \end{pmatrix}$ ,  $\begin{pmatrix} 1+2i \\ -3 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 1-i \\ -i \\ 1 \end{pmatrix}$ , (g)  $\begin{pmatrix} 1+i \\ 2-i \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 1-i \\ -3i \\ 1-2i \end{pmatrix}$ ,  $\begin{pmatrix} -1+i \\ 2+3i \\ 1+2i \end{pmatrix}$ .

3.6.27. *True or false:* The set of complex vectors of the form  $\begin{pmatrix} z \\ \bar{z} \end{pmatrix}$  for  $z \in \mathbb{C}$  is a subspace of  $\mathbb{C}^2$ .

3.6.28. (a) Determine whether the vectors  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1+i \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} -1+i \\ 1+i \\ -1 \end{pmatrix}$ ,

are linearly independent or linearly dependent. (b) Do they form a basis of  $\mathbb{C}^3$ ?

(c) Compute the Hermitian norm of each vector. (d) Compute the Hermitian dot products between all different pairs. Which vectors are orthogonal?

3.6.29. Find the dimension of and a basis for the following subspaces of  $\mathbb{C}^3$ : (a) The set of all complex multiples of  $(1, i, 1-i)^T$ . (b) The plane  $z_1 + iz_2 + (1-i)z_3 = 0$ . (c) The image of the matrix  $A = \begin{pmatrix} 1 & i & 2-i \\ 2+i & 1+3i & -1-i \end{pmatrix}$ . (d) The kernel of the same matrix. (e) The set of vectors that are orthogonal to  $(1-i, 2i, 1+i)^T$ .

3.6.30. Find bases for the four fundamental subspaces associated with the complex matrices

(a)  $\begin{pmatrix} i & 2 \\ -1 & 2i \end{pmatrix}$ , (b)  $\begin{pmatrix} 2 & -1+i & 1-2i \\ -4 & 3-i & 1+i \end{pmatrix}$ , (c)  $\begin{pmatrix} i & -1 & 2-i \\ -1+2i & -2-i & 3 \\ i & -1 & 1+i \end{pmatrix}$ .

3.6.31. Prove that  $\mathbf{v} = \mathbf{x} + i\mathbf{y}$  and  $\bar{\mathbf{v}} = \mathbf{x} - i\mathbf{y}$  are linearly independent complex vectors if and only if their real and imaginary parts  $\mathbf{x}$  and  $\mathbf{y}$  are linearly independent real vectors.

3.6.32. Prove that the space of complex  $m \times n$  matrices is a complex vector space. What is its dimension?

3.6.33. Determine which of the following are subspaces of the vector space consisting of all complex  $2 \times 2$  matrices. (a) All matrices with real diagonals. (b) All matrices for which the sum of the diagonal entries is zero. (c) All singular complex matrices. (d) All matrices whose determinant is real. (e) All matrices of the form  $\begin{pmatrix} a & b \\ \bar{a} & \bar{b} \end{pmatrix}$ , where  $a, b \in \mathbb{C}$ .

3.6.34. *True or false:* The set of all complex-valued functions  $u(x) = v(x) + i w(x)$  with  $u(0) = i$  is a subspace of the vector space of complex-valued functions.

3.6.35. Let  $V$  denote the complex vector space spanned by the functions  $1, e^{ix}$  and  $e^{-ix}$ , where  $x$  is a real variable. Which of the following functions belong to  $V$ ?

- (a)  $\sin x$ , (b)  $\cos x - 2i \sin x$ , (c)  $\cosh x$ , (d)  $\sin^2 \frac{1}{2}x$ , (e)  $\cos^2 x$ ?

3.6.36. Prove that the following define Hermitian inner products on  $\mathbb{C}^2$ :

- (a)  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 \bar{w}_1 + 2v_2 \bar{w}_2$ , (b)  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 \bar{w}_1 + i v_1 \bar{w}_2 - i v_2 \bar{w}_1 + 2v_2 \bar{w}_2$ .

3.6.37. Which of the following define inner products on  $\mathbb{C}^2$ ? (a)  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 \bar{w}_1 + 2i v_2 \bar{w}_2$ ,

- (b)  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2$ , (c)  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 \bar{w}_2 + v_2 \bar{w}_1$ , (d)  $\langle \mathbf{v}, \mathbf{w} \rangle =$

- $2v_1 \bar{w}_1 + v_1 \bar{w}_2 + v_2 \bar{w}_1 + 2v_2 \bar{w}_2$ , (e)  $\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1 \bar{w}_1 + (1+i)v_1 \bar{w}_2 + (1-i)v_2 \bar{w}_1 + 3v_2 \bar{w}_2$ .

◊ 3.6.38. Let  $A = A^T$  be a real symmetric  $n \times n$  matrix. Show that  $(A\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\mathbf{w})$  for all  $\mathbf{v}, \mathbf{w} \in \mathbb{C}^n$ .

3.6.39. Let  $\mathbf{z} = \mathbf{x} + i\mathbf{y} \in \mathbb{C}^n$ .

- (a) Prove that, for the Hermitian dot product,  $\|\mathbf{z}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ .

- (b) Does this formula remain valid under a more general Hermitian inner product on  $\mathbb{C}^n$ ?

◊ 3.6.40. Let  $V$  be a complex inner product space. Prove that, for all  $\mathbf{z}, \mathbf{w} \in V$ ,

- (a)  $\|\mathbf{z} + \mathbf{w}\|^2 = \|\mathbf{z}\|^2 + 2\operatorname{Re} \langle \mathbf{z}, \mathbf{w} \rangle + \|\mathbf{w}\|^2$ ;

- (b)  $\langle \mathbf{z}, \mathbf{w} \rangle = \frac{1}{4} (\|\mathbf{z} + \mathbf{w}\|^2 - \|\mathbf{z} - \mathbf{w}\|^2 + i\|\mathbf{z} + i\mathbf{w}\|^2 - i\|\mathbf{z} - i\mathbf{w}\|^2)$ .

◊ 3.6.41. (a) How would you define the angle between two elements of a complex inner product space? (b) What is the angle between  $(-1, 2 - i, -1 + 2i)^T$  and  $(-2 - i, -i, 1 - i)^T$  relative to the Hermitian dot product?

3.6.42. Let  $\mathbf{0} \neq \mathbf{v} \in \mathbb{C}^n$ . Which scalar multiples  $c\mathbf{v}$  have the same Hermitian norm as  $\mathbf{v}$ ?

◊ 3.6.43. Prove the Cauchy–Schwarz inequality (3.103) and the triangle inequality (3.104) for a general complex inner product. *Hint:* Use Exercises 3.6.8, 3.6.40(a).

◊ 3.6.44. The *Hermitian adjoint* of a complex  $m \times n$  matrix  $A$  is the complex conjugate of its transpose, written  $A^\dagger = \overline{A^T} = \overline{A}^T$ .

For example, if  $A = \begin{pmatrix} 1+i & 2i \\ -3 & 2-5i \end{pmatrix}$ , then  $A^\dagger = \begin{pmatrix} 1-i & -3 \\ -2i & 2+5i \end{pmatrix}$ . Prove that

- (a)  $(A^\dagger)^\dagger = A$ , (b)  $(zA + wB)^\dagger = \bar{z}A^\dagger + \bar{w}B^\dagger$  for  $z, w \in \mathbb{C}$ , (c)  $(AB)^\dagger = B^\dagger A^\dagger$ .

◊ 3.6.45. A complex matrix  $H$  is called *Hermitian* if it equals its Hermitian adjoint,  $H^\dagger = H$ , as defined in the preceding exercise. (a) Prove that the diagonal entries of a Hermitian matrix are real. (b) Prove that  $(H\mathbf{z}) \cdot \mathbf{w} = \mathbf{z} \cdot (H\mathbf{w})$  for  $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$ . (c) Prove that every Hermitian inner product on  $\mathbb{C}^n$  has the form  $\langle \mathbf{z}, \mathbf{w} \rangle = \mathbf{z}^T H \bar{\mathbf{w}}$ , where  $H$  is an  $n \times n$  positive definite Hermitian matrix. (d) How would you verify positive definiteness of a complex matrix?

3.6.46. *Multiple choice:* Let  $V$  be a complex normed vector space. How many unit vectors are parallel to a given vector  $\mathbf{0} \neq \mathbf{v} \in V$ ? (a) none; (b) 1; (c) 2; (d) 3; (e)  $\infty$ ; (f) depends upon the vector; (g) depends on the norm. Explain your answer.

◊ 3.6.47. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be elements of a complex inner product space. Let  $K$  denote the corresponding  $n \times n$  *Gram matrix*, defined in the usual manner.

- (a) Prove that  $K$  is a Hermitian matrix, as defined in Exercise 3.6.45.
- (b) Prove that  $K$  is positive semi-definite, meaning  $\mathbf{z}^T K \mathbf{z} \geq 0$  for all  $\mathbf{z} \in \mathbb{C}^n$ .
- (c) Prove that  $K$  is positive definite if and only if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are linearly independent.

3.6.48. For each of the following pairs of complex-valued functions,

- (i) compute their  $L^2$  norm and Hermitian inner product on the interval  $[0, 1]$ , and then
- (ii) check the validity of the Cauchy–Schwarz and triangle inequalities.

$$(a) 1, e^{i\pi x}; \quad (b) x + i, x - i; \quad (c) i x^2, (1 - 2i)x + 3i.$$

3.6.49. Formulate conditions on a weight function  $w(x)$  that guarantee that the weighted integral  $\langle f, g \rangle = \int_a^b f(x) \overline{g(x)} w(x) dx$  defines an inner product on the space of continuous complex-valued functions on  $[a, b]$ .

3.6.50. (a) Formulate a general definition of a norm on a complex vector space.

(b) How would you define analogues of the  $L^1, L^2$  and  $L^\infty$  norms on  $\mathbb{C}^n$ ?

---



# Chapter 4

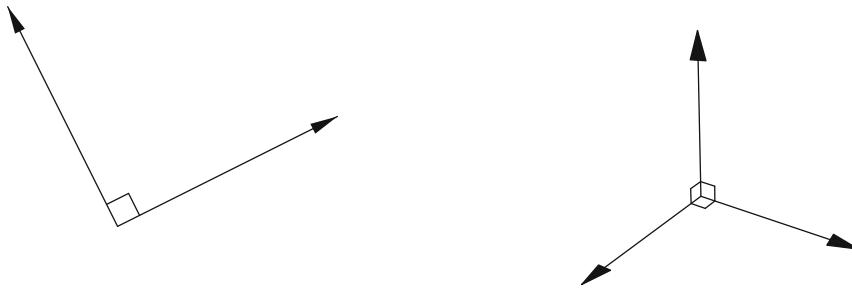
## Orthogonality

Orthogonality is the mathematical formalization of the geometrical property of perpendicularity, as adapted to general inner product spaces. In linear algebra, bases consisting of mutually orthogonal elements play an essential role in theoretical developments, in a broad range of applications, and in the design of practical numerical algorithms. Computations become dramatically simpler and less prone to numerical instabilities when performed in orthogonal coordinate systems. Indeed, many large-scale modern applications would be impractical, if not completely infeasible, were it not for the dramatic simplifying power of orthogonality.

The duly famous Gram–Schmidt process will convert an arbitrary basis of an inner product space into an orthogonal basis. In Euclidean space, the Gram–Schmidt process can be reinterpreted as a new kind of matrix factorization, in which a nonsingular matrix  $A = QR$  is written as the product of an orthogonal matrix  $Q$  and an upper triangular matrix  $R$ . The  $QR$  factorization and its generalizations are used in statistical data analysis as well as the design of numerical algorithms for computing eigenvalues and eigenvectors. In function space, the Gram–Schmidt algorithm is employed to construct orthogonal polynomials and other useful systems of orthogonal functions.

Orthogonality is motivated by geometry, and orthogonal matrices, meaning those whose columns form an orthonormal system, are of fundamental importance in the mathematics of symmetry, in image processing, and in computer graphics, animation, and cinema, [5, 12, 72, 73]. The orthogonal projection of a point onto a subspace turns out to be the closest point or least squares minimizer, as we discuss in Chapter 5. Yet another important fact is that the four fundamental subspaces of a matrix that were introduced in Chapter 2 come in mutually orthogonal pairs. This observation leads directly to a new characterization of the compatibility conditions for linear algebraic systems known as the Fredholm alternative, whose extensions are used in the analysis of linear boundary value problems, differential equations, and integral equations, [16, 61]. The orthogonality of eigenvector and eigenfunction bases for symmetric matrices and self-adjoint operators provides the key to understanding the dynamics of discrete and continuous mechanical, thermodynamical, electrical, and quantum mechanical systems.

One of the most fertile applications of orthogonal bases is in signal processing. Fourier analysis decomposes a signal into its simple periodic components — sines and cosines — which form an orthogonal system of functions, [61, 77]. Modern digital media, such as CD's, DVD's and MP3's, are based on discrete data obtained by sampling a physical signal. The Discrete Fourier Transform (DFT) uses orthogonality to decompose the sampled signal vector into a linear combination of sampled trigonometric functions (or, more accurately, complex exponentials). Basic data compression and noise removal algorithms are applied to the discrete Fourier coefficients, acting on the observation that noise tends to accumulate in the high-frequency Fourier modes. More sophisticated signal and image processing techniques, including smoothing and compression algorithms, are based on orthogonal wavelet bases, which are discussed in Section 9.7.



**Figure 4.1.** Orthonormal Bases in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ .

## 4.1 Orthogonal and Orthonormal Bases

Let  $V$  be a real<sup>†</sup> inner product space. Recall that two elements  $\mathbf{v}, \mathbf{w} \in V$  are called *orthogonal* if their inner product vanishes:  $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ . In the case of vectors in Euclidean space, orthogonality under the dot product means that they meet at a right angle.

A particularly important configuration arises when  $V$  admits a basis consisting of mutually orthogonal elements.

**Definition 4.1.** A basis  $\mathbf{u}_1, \dots, \mathbf{u}_n$  of an  $n$ -dimensional inner product space  $V$  is called *orthogonal* if  $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$  for all  $i \neq j$ . The basis is called *orthonormal* if, in addition, each vector has unit length:  $\|\mathbf{u}_i\| = 1$ , for all  $i = 1, \dots, n$ .

For the Euclidean space  $\mathbb{R}^n$  equipped with the standard dot product, the simplest example of an orthonormal basis is the standard basis

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \dots \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Orthogonality follows because  $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ , for  $i \neq j$ , while  $\|\mathbf{e}_i\| = 1$  implies normality.

Since a basis cannot contain the zero vector, there is an easy way to convert an orthogonal basis to an orthonormal basis. Namely, we replace each basis vector with a unit vector pointing in the same direction, as in Lemma 3.14.

**Lemma 4.2.** If  $\mathbf{v}_1, \dots, \mathbf{v}_n$  is an orthogonal basis of a vector space  $V$ , then the normalized vectors  $\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$ ,  $i = 1, \dots, n$ , form an orthonormal basis.

<sup>†</sup> The methods can be adapted more or less straightforwardly to the complex realm. The main complication, as noted in Section 3.6, is that we need to be careful with the order of vectors appearing in the conjugate symmetric complex inner products. In this chapter, we will be careful to write the inner product formulas in the proper order so that they retain their validity in complex vector spaces.

**Example 4.3.** The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix},$$

are easily seen to form a basis of  $\mathbb{R}^3$ . Moreover, they are mutually perpendicular,  $\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_1 \cdot \mathbf{v}_3 = \mathbf{v}_2 \cdot \mathbf{v}_3 = 0$ , and so form an orthogonal basis with respect to the standard dot product on  $\mathbb{R}^3$ . When we divide each orthogonal basis vector by its length, the result is the orthonormal basis

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{30}} \begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{\sqrt{30}} \\ -\frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{pmatrix},$$

satisfying  $\mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$  and  $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = 1$ . The appearance of square roots in the elements of an orthonormal basis is fairly typical.

A useful observation is that every orthogonal collection of nonzero vectors is automatically linearly independent.

**Proposition 4.4.** Let  $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$  be nonzero, mutually orthogonal elements, so  $\mathbf{v}_i \neq \mathbf{0}$  and  $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$  for all  $i \neq j$ . Then  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are linearly independent.

*Proof:* Suppose

$$c_1 \mathbf{v}_1 + \dots + c_k \mathbf{v}_k = \mathbf{0}.$$

Let us take the inner product of this equation with any  $\mathbf{v}_i$ . Using linearity of the inner product and orthogonality, we compute

$$0 = \langle c_1 \mathbf{v}_1 + \dots + c_k \mathbf{v}_k, \mathbf{v}_i \rangle = c_1 \langle \mathbf{v}_1, \mathbf{v}_i \rangle + \dots + c_k \langle \mathbf{v}_k, \mathbf{v}_i \rangle = c_i \langle \mathbf{v}_i, \mathbf{v}_i \rangle = c_i \|\mathbf{v}_i\|^2.$$

Therefore, given that  $\mathbf{v}_i \neq \mathbf{0}$ , we conclude that  $c_i = 0$ . Since this holds for all  $i = 1, \dots, k$ , the linear independence of  $\mathbf{v}_1, \dots, \mathbf{v}_k$  follows. *Q.E.D.*

As a direct corollary, we infer that every collection of nonzero orthogonal vectors forms a basis for its span.

**Theorem 4.5.** Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$  are nonzero, mutually orthogonal elements of an inner product space  $V$ . Then  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form an orthogonal basis for their span  $W = \text{span} \{ \mathbf{v}_1, \dots, \mathbf{v}_n \} \subset V$ , which is therefore a subspace of dimension  $n = \dim W$ . In particular, if  $\dim V = n$ , then  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a orthogonal basis for  $V$ .

Orthogonality is also of profound significance for function spaces. Here is a relatively simple example.

**Example 4.6.** Consider the vector space  $\mathcal{P}^{(2)}$  consisting of all quadratic polynomials  $p(x) = \alpha + \beta x + \gamma x^2$ , equipped with the  $L^2$  inner product and norm

$$\langle p, q \rangle = \int_0^1 p(x) q(x) dx, \quad \|p\| = \sqrt{\langle p, p \rangle} = \sqrt{\int_0^1 p(x)^2 dx}.$$

The standard monomials  $1, x, x^2$  do *not* form an orthogonal basis. Indeed,

$$\langle 1, x \rangle = \frac{1}{2}, \quad \langle 1, x^2 \rangle = \frac{1}{3}, \quad \langle x, x^2 \rangle = \frac{1}{4}.$$

One orthogonal basis of  $\mathcal{P}^{(2)}$  is provided by following polynomials:

$$p_1(x) = 1, \quad p_2(x) = x - \frac{1}{2}, \quad p_3(x) = x^2 - x + \frac{1}{6}. \quad (4.1)$$

Indeed, one easily verifies that  $\langle p_1, p_2 \rangle = \langle p_1, p_3 \rangle = \langle p_2, p_3 \rangle = 0$ , while

$$\|p_1\| = 1, \quad \|p_2\| = \frac{1}{\sqrt{12}} = \frac{1}{2\sqrt{3}}, \quad \|p_3\| = \frac{1}{\sqrt{180}} = \frac{1}{6\sqrt{5}}.$$

The corresponding orthonormal basis is found by dividing each orthogonal basis element by its norm:

$$u_1(x) = 1, \quad u_2(x) = \sqrt{3} (2x - 1), \quad u_3(x) = \sqrt{5} (6x^2 - 6x + 1).$$

In Section 4.5 below, we will learn how to systematically construct such orthogonal systems of polynomials.

## Exercises

4.1.1. Let  $\mathbb{R}^2$  have the standard dot product. Classify the following pairs of vectors as

(i) basis, (ii) orthogonal basis, and/or (iii) orthonormal basis:

$$(a) \mathbf{v}_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}; (b) \mathbf{v}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}; (c) \mathbf{v}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix};$$

$$(d) \mathbf{v}_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ -6 \end{pmatrix}; (e) \mathbf{v}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}; (f) \mathbf{v}_1 = \begin{pmatrix} \frac{3}{5} \\ \frac{4}{5} \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -\frac{4}{5} \\ \frac{3}{5} \end{pmatrix}.$$

4.1.2. Let  $\mathbb{R}^3$  have the standard dot product. Classify the following sets of vectors as

(i) basis, (ii) orthogonal basis, and/or (iii) orthonormal basis:

$$(a) \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}; (b) \begin{pmatrix} -\frac{4}{13} \\ \frac{3}{5} \\ -\frac{48}{65} \end{pmatrix}, \begin{pmatrix} \frac{12}{13} \\ 0 \\ -\frac{5}{13} \end{pmatrix}, \begin{pmatrix} \frac{3}{13} \\ \frac{4}{5} \\ \frac{36}{65} \end{pmatrix}; (c) \begin{pmatrix} 0 \\ -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}.$$

4.1.3. Repeat Exercise 4.1.1, but use the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + \frac{1}{9} v_2 w_2$  instead of the dot product.

4.1.4. Show that the standard basis vectors  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  form an orthogonal basis with respect to the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$  on  $\mathbb{R}^3$ . Find an orthonormal basis for this inner product space.

4.1.5. Find all values of  $a$  such that the vectors  $\begin{pmatrix} a \\ 1 \end{pmatrix}, \begin{pmatrix} -a \\ 1 \end{pmatrix}$  form an orthogonal basis of

$\mathbb{R}^2$  under (a) the dot product; (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 3v_1 w_1 + 2v_2 w_2$ ;  
(c) the inner product prescribed by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ .

4.1.6. Find all possible values of  $a$  and  $b$  in the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = a v_1 w_1 + b v_2 w_2$  that make the vectors  $(1, 2)^T, (-1, 1)^T$ , an orthogonal basis in  $\mathbb{R}^2$ .

4.1.7. Answer Exercise 4.1.6 for the vectors (a)  $(2, 3)^T, (-2, 2)^T$ ; (b)  $(1, 4)^T, (2, 1)^T$ .

4.1.8. Find an inner product such that the vectors  $(-1, 2)^T$  and  $(1, 2)^T$  form an orthonormal basis of  $\mathbb{R}^2$ .

4.1.9. *True or false:* If  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  are a basis for  $\mathbb{R}^3$ , then they form an orthogonal basis under some appropriately weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = a v_1 w_1 + b v_2 w_2 + c v_3 w_3$ .

◇ 4.1.10. The *cross product* between two vectors in  $\mathbb{R}^3$  is the vector defined by the formula

$$\mathbf{v} \times \mathbf{w} = \begin{pmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{pmatrix}, \quad \text{where} \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}. \quad (4.2)$$

- (a) Show that  $\mathbf{u} = \mathbf{v} \times \mathbf{w}$  is orthogonal, under the dot product, to both  $\mathbf{v}$  and  $\mathbf{w}$ .
- (b) Show that  $\mathbf{v} \times \mathbf{w} = \mathbf{0}$  if and only if  $\mathbf{v}$  and  $\mathbf{w}$  are parallel. (c) Prove that if  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$  are orthogonal nonzero vectors, then  $\mathbf{u} = \mathbf{v} \times \mathbf{w}$ ,  $\mathbf{v}, \mathbf{w}$  form an orthogonal basis of  $\mathbb{R}^3$ .
- (d) *True or false:* If  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$  are orthogonal unit vectors, then  $\mathbf{v}, \mathbf{w}$  and  $\mathbf{u} = \mathbf{v} \times \mathbf{w}$  form an orthonormal basis of  $\mathbb{R}^3$ .

◇ 4.1.11. Prove that every orthonormal basis of  $\mathbb{R}^2$  under the standard dot product has the form  $\mathbf{u}_1 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}$  and  $\mathbf{u}_2 = \pm \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}$  for some  $0 \leq \theta < 2\pi$  and some choice of  $\pm$  sign.

◇ 4.1.12. Given angles  $\theta, \varphi, \psi$ , prove that the vectors  $\mathbf{u}_1 = \begin{pmatrix} \cos \psi \cos \varphi - \cos \theta \sin \varphi \sin \psi \\ -\sin \psi \cos \varphi - \cos \theta \sin \varphi \cos \psi \\ \sin \theta \sin \varphi \end{pmatrix}$ ,

$$\mathbf{u}_2 = \begin{pmatrix} \cos \psi \sin \varphi + \cos \theta \cos \varphi \sin \psi \\ -\sin \psi \sin \varphi + \cos \theta \cos \varphi \cos \psi \\ -\sin \theta \cos \varphi \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \sin \theta \sin \varphi \\ \sin \theta \cos \varphi \\ \cos \theta \end{pmatrix}, \quad \text{form an orthonormal basis}$$

of  $\mathbb{R}^3$  under the standard dot product. **Remark.** It can be proved, [31; p. 147], that *every* orthonormal basis of  $\mathbb{R}^3$  has the form  $\mathbf{u}_1, \mathbf{u}_2, \pm \mathbf{u}_3$  for some choice of angles  $\theta, \varphi, \psi$ .

◇ 4.1.13. (a) Show that  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form an orthonormal basis of  $\mathbb{R}^n$  for the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T K \mathbf{w}$  for  $K > 0$  if and only if  $A^T K A = I$ , where  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ .  
 (b) Prove that every basis of  $\mathbb{R}^n$  is an orthonormal basis with respect to some inner product. Is the inner product uniquely determined? (c) Find the inner product on  $\mathbb{R}^2$  that makes  $\mathbf{v}_1 = (1, 1)^T, \mathbf{v}_2 = (2, 3)^T$  into an orthonormal basis. (d) Find the inner product on  $\mathbb{R}^3$  that makes  $\mathbf{v}_1 = (1, 1, 1)^T, \mathbf{v}_2 = (1, 1, 2)^T, \mathbf{v}_3 = (1, 2, 3)^T$  an orthonormal basis.

4.1.14. Describe all orthonormal bases of  $\mathbb{R}^2$  for the inner products

$$(a) \langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \mathbf{w}; \quad (b) \langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \mathbf{w}.$$

4.1.15. Let  $\mathbf{v}$  and  $\mathbf{w}$  be elements of an inner product space. Prove that

$\|\mathbf{v} + \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2$  if and only if  $\mathbf{v}, \mathbf{w}$  are orthogonal. Explain why this formula can be viewed as the generalization of the Pythagorean Theorem.

4.1.16. Prove that if  $\mathbf{v}_1, \mathbf{v}_2$  form a basis of an inner product space  $V$  and  $\|\mathbf{v}_1\| = \|\mathbf{v}_2\|$ , then  $\mathbf{v}_1 + \mathbf{v}_2$  and  $\mathbf{v}_1 - \mathbf{v}_2$  form an orthogonal basis of  $V$ .

4.1.17. Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are nonzero mutually orthogonal elements of an inner product space  $V$ . Write down their Gram matrix. Why is it nonsingular?

4.1.18. Let  $V = \mathcal{P}^{(1)}$  be the vector space consisting of linear polynomials  $p(t) = at + b$ .

- (a) Carefully explain why  $\langle p, q \rangle = \int_0^1 t p(t) q(t) dt$  defines an inner product on  $V$ .
- (b) Find all polynomials  $p(t) = at + b \in V$  that are orthogonal to  $p_1(t) = 1$  based on this inner product. (c) Use part (b) to construct an orthonormal basis of  $V$  for this inner product. (d) Find an orthonormal basis of the space  $\mathcal{P}^{(2)}$  of quadratic polynomials for the same inner product. *Hint:* First find a quadratic polynomial that is orthogonal to the basis you constructed in part (c).

4.1.19. Explain why the functions  $\cos x, \sin x$  form an orthogonal basis for the space of solutions to the differential equation  $y'' + y = 0$  under the  $L^2$  inner product on  $[-\pi, \pi]$ .

4.1.20. Do the functions  $e^{x/2}, e^{-x/2}$  form an orthogonal basis for the space of solutions to the differential equation  $4y'' - y = 0$  under the  $L^2$  inner product on  $[0, 1]$ ? If not, can you find an orthogonal basis of the solution space?

## Computations in Orthogonal Bases

What are the advantages of orthogonal and orthonormal bases? Once one has a basis of a vector space, a key issue is how to express other elements as linear combinations of the basis elements — that is, to find their *coordinates* in the prescribed basis. In general, this is not so easy, since it requires solving a system of linear equations, as described in (2.23). In high-dimensional situations arising in applications, computing the solution may require a considerable, if not infeasible, amount of time and effort.

However, if the basis is orthogonal, or, even better, orthonormal, then the change of basis computation requires almost no work. This is the crucial insight underlying the efficacy of both discrete and continuous Fourier analysis in signal, image, and video processing, least squares approximations, the statistical analysis of large data sets, and a multitude of other applications, both classical and modern.

**Theorem 4.7.** Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be an orthonormal basis for an inner product space  $V$ . Then one can write any element  $\mathbf{v} \in V$  as a linear combination

$$\mathbf{v} = c_1 \mathbf{u}_1 + \dots + c_n \mathbf{u}_n, \quad (4.3)$$

in which its *coordinates*

$$c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle, \quad i = 1, \dots, n, \quad (4.4)$$

are explicitly given as inner products. Moreover, its norm is given by the *Pythagorean formula*

$$\|\mathbf{v}\| = \sqrt{c_1^2 + \dots + c_n^2} = \sqrt{\sum_{i=1}^n \langle \mathbf{v}, \mathbf{u}_i \rangle^2}, \quad (4.5)$$

namely, the square root of the sum of the squares of its orthonormal basis coordinates.

*Proof:* Let us compute the inner product of the element (4.3) with one of the basis vectors. Using the orthonormality conditions

$$\langle \mathbf{u}_i, \mathbf{u}_j \rangle = \begin{cases} 0 & i \neq j, \\ 1 & i = j, \end{cases} \quad (4.6)$$

and bilinearity of the inner product, we obtain

$$\langle \mathbf{v}, \mathbf{u}_i \rangle = \left\langle \sum_{j=1}^n c_j \mathbf{u}_j, \mathbf{u}_i \right\rangle = \sum_{j=1}^n c_j \langle \mathbf{u}_j, \mathbf{u}_i \rangle = c_i \|\mathbf{u}_i\|^2 = c_i.$$

To prove formula (4.5), we similarly expand

$$\|\mathbf{v}\|^2 = \langle \mathbf{v}, \mathbf{v} \rangle = \left\langle \sum_{j=1}^n c_j \mathbf{u}_j, \sum_{j=1}^n c_j \mathbf{u}_j \right\rangle = \sum_{i,j=1}^n c_i c_j \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \sum_{i=1}^n c_i^2,$$

again making use of the orthonormality of the basis elements.

*Q.E.D.*

It is worth emphasizing that the Pythagorean-type formula (4.5) is valid for *all* inner products.

**Example 4.8.** Let us rewrite the vector  $\mathbf{v} = (1, 1, 1)^T$  in terms of the orthonormal basis

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{5}{\sqrt{30}} \\ -\frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{pmatrix},$$

constructed in Example 4.3. Computing the dot products

$$\mathbf{v} \cdot \mathbf{u}_1 = \frac{2}{\sqrt{6}}, \quad \mathbf{v} \cdot \mathbf{u}_2 = \frac{3}{\sqrt{5}}, \quad \mathbf{v} \cdot \mathbf{u}_3 = \frac{4}{\sqrt{30}},$$

we immediately conclude that

$$\mathbf{v} = \frac{2}{\sqrt{6}} \mathbf{u}_1 + \frac{3}{\sqrt{5}} \mathbf{u}_2 + \frac{4}{\sqrt{30}} \mathbf{u}_3.$$

Needless to say, a direct computation based on solving the associated linear system, as in Chapter 2, is more tedious.

While passage from an orthogonal basis to its orthonormal version is elementary — one simply divides each basis element by its norm — we shall often find it more convenient to work directly with the unnormalized version. The next result provides the corresponding formula expressing a vector in terms of an orthogonal, but not necessarily orthonormal basis. The proof proceeds exactly as in the orthonormal case, and details are left to the reader.

**Theorem 4.9.** If  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form an orthogonal basis, then the corresponding coordinates of a vector

$$\mathbf{v} = a_1 \mathbf{v}_1 + \dots + a_n \mathbf{v}_n \quad \text{are given by} \quad a_i = \frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\| \mathbf{v}_i \|^2}. \quad (4.7)$$

In this case, its norm can be computed using the formula

$$\| \mathbf{v} \|^2 = \sum_{i=1}^n a_i^2 \| \mathbf{v}_i \|^2 = \sum_{i=1}^n \left( \frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\| \mathbf{v}_i \|^2} \right)^2. \quad (4.8)$$

Equation (4.7), along with its orthonormal simplification (4.4), is one of the most useful formulas we shall establish, and applications will appear repeatedly throughout this text and beyond.

**Example 4.10.** The wavelet basis

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \quad (4.9)$$

introduced in Example 2.35 is, in fact, an orthogonal basis of  $\mathbb{R}^4$ . The norms are

$$\| \mathbf{v}_1 \| = 2, \quad \| \mathbf{v}_2 \| = 2, \quad \| \mathbf{v}_3 \| = \sqrt{2}, \quad \| \mathbf{v}_4 \| = \sqrt{2}.$$

Therefore, using (4.7), we can readily express any vector as a linear combination of the wavelet basis vectors. For example,

$$\mathbf{v} = \begin{pmatrix} 4 \\ -2 \\ 1 \\ 5 \end{pmatrix} = 2\mathbf{v}_1 - \mathbf{v}_2 + 3\mathbf{v}_3 - 2\mathbf{v}_4,$$

where the wavelet coordinates are computed directly by

$$\frac{\langle \mathbf{v}, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} = \frac{8}{4} = 2, \quad \frac{\langle \mathbf{v}, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} = \frac{-4}{4} = -1, \quad \frac{\langle \mathbf{v}, \mathbf{v}_3 \rangle}{\|\mathbf{v}_3\|^2} = \frac{6}{2} = 3, \quad \frac{\langle \mathbf{v}, \mathbf{v}_4 \rangle}{\|\mathbf{v}_4\|^2} = \frac{-4}{2} = -2.$$

This is clearly quicker than solving the linear system, as we did earlier in Example 2.35. Finally, we note that

$$46 = \|\mathbf{v}\|^2 = 2^2\|\mathbf{v}_1\|^2 + (-1)^2\|\mathbf{v}_2\|^2 + 3^2\|\mathbf{v}_3\|^2 + (-2)^2\|\mathbf{v}_4\|^2 = 4 \cdot 4 + 1 \cdot 4 + 9 \cdot 2 + 4 \cdot 2,$$

in conformity with (4.8).

**Example 4.11.** The same formulas are equally valid for orthogonal bases in function spaces. For example, to express a quadratic polynomial

$$p(x) = c_1 p_1(x) + c_2 p_2(x) + c_3 p_3(x) = c_1 + c_2 \left( x - \frac{1}{2} \right) + c_3 \left( x^2 - x + \frac{1}{6} \right)$$

in terms of the orthogonal basis (4.1), we merely compute the  $L^2$  inner product integrals

$$\begin{aligned} c_1 &= \frac{\langle p, p_1 \rangle}{\|p_1\|^2} = \int_0^1 p(x) dx, & c_2 &= \frac{\langle p, p_2 \rangle}{\|p_2\|^2} = 12 \int_0^1 p(x) \left( x - \frac{1}{2} \right) dx, \\ c_3 &= \frac{\langle p, p_3 \rangle}{\|p_3\|^2} = 180 \int_0^1 p(x) \left( x^2 - x + \frac{1}{6} \right) dx. \end{aligned}$$

Thus, for example, the coefficients for  $p(x) = x^2 + x + 1$  are

$$\begin{aligned} c_1 &= \int_0^1 (x^2 + x + 1) dx = \frac{11}{6}, & c_2 &= 12 \int_0^1 (x^2 + x + 1) \left( x - \frac{1}{2} \right) dx = 2, \\ c_3 &= 180 \int_0^1 (x^2 + x + 1) \left( x^2 - x + \frac{1}{6} \right) dx = 1, \end{aligned}$$

and so

$$p(x) = x^2 + x + 1 = \frac{11}{6} + 2 \left( x - \frac{1}{2} \right) + \left( x^2 - x + \frac{1}{6} \right).$$

**Example 4.12.** Perhaps the most important example of an orthogonal basis is provided by the basic trigonometric functions. Let  $\mathcal{T}^{(n)}$  denote the vector space consisting of all *trigonometric polynomials*

$$T(x) = \sum_{0 \leq j+k \leq n} a_{jk} (\sin x)^j (\cos x)^k \quad (4.10)$$

of *degree*  $\leq n$ . The individual monomials  $(\sin x)^j (\cos x)^k$  span  $\mathcal{T}^{(n)}$ , but, as we saw in Example 2.20, they do not form a basis, owing to identities stemming from the basic

trigonometric formula  $\cos^2 x + \sin^2 x = 1$ . Exercise 3.6.21 introduced a more convenient spanning set consisting of the  $2n + 1$  functions

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x, \quad \dots \quad \cos nx, \quad \sin nx. \quad (4.11)$$

Let us prove that these functions form an orthogonal basis of  $\mathcal{T}^{(n)}$  with respect to the  $L^2$  inner product and norm:

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) g(x) dx, \quad \|f\|^2 = \int_{-\pi}^{\pi} f(x)^2 dx. \quad (4.12)$$

The elementary integration formulas

$$\int_{-\pi}^{\pi} \cos kx \cos lx dx = \begin{cases} 0, & k \neq l, \\ 2\pi, & k = l = 0, \\ \pi, & k = l \neq 0, \end{cases} \quad \int_{-\pi}^{\pi} \sin kx \sin lx dx = \begin{cases} 0, & k \neq l, \\ \pi, & k = l \neq 0, \end{cases} \quad \int_{-\pi}^{\pi} \cos kx \sin lx dx = 0, \quad (4.13)$$

which are valid for all nonnegative integers  $k, l \geq 0$ , imply the orthogonality relations

$$\begin{aligned} \langle \cos kx, \cos lx \rangle &= \langle \sin kx, \sin lx \rangle = 0, & k \neq l, \quad \langle \cos kx, \sin lx \rangle &= 0, \\ \|\cos kx\| &= \|\sin kx\| = \sqrt{\pi}, & k \neq 0, \quad \|1\| &= \sqrt{2\pi}. \end{aligned} \quad (4.14)$$

Theorem 4.5 now assures us that the functions (4.11) form a basis for  $\mathcal{T}^{(n)}$ . One consequence is that  $\dim \mathcal{T}^{(n)} = 2n + 1$  — a fact that is not so easy to establish directly.

Orthogonality of the trigonometric functions (4.11) means that we can compute the coefficients  $a_0, \dots, a_n, b_1, \dots, b_n$  of any trigonometric polynomial

$$p(x) = a_0 + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx) \quad (4.15)$$

by an explicit integration formula. Namely,

$$\begin{aligned} a_0 &= \frac{\langle f, 1 \rangle}{\|1\|^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx, & a_k &= \frac{\langle f, \cos kx \rangle}{\|\cos kx\|^2} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, \\ b_k &= \frac{\langle f, \sin kx \rangle}{\|\sin kx\|^2} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx dx, & k \geq 1. \end{aligned} \quad (4.16)$$

These fundamental formulas play an essential role in the theory and applications of Fourier series, [61, 79, 77].

## Exercises

4.1.21. (a) Prove that the vectors  $\mathbf{v}_1 = (1, 1, 1)^T$ ,  $\mathbf{v}_2 = (1, 1, -2)^T$ ,  $\mathbf{v}_3 = (-1, 1, 0)^T$ , form an orthogonal basis of  $\mathbb{R}^3$  with the dot product. (b) Use orthogonality to write the vector  $\mathbf{v} = (1, 2, 3)^T$  as a linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ . (c) Verify the formula (4.8) for  $\|\mathbf{v}\|$ . (d) Construct an orthonormal basis, using the given vectors. (e) Write  $\mathbf{v}$  as a linear combination of the orthonormal basis, and verify (4.5).

4.1.22. (a) Prove that  $\mathbf{v}_1 = \left(\frac{3}{5}, 0, \frac{4}{5}\right)^T$ ,  $\mathbf{v}_2 = \left(-\frac{4}{13}, \frac{12}{13}, \frac{3}{13}\right)^T$ ,  $\mathbf{v}_3 = \left(-\frac{48}{65}, -\frac{5}{13}, \frac{36}{65}\right)^T$ , form an orthonormal basis for  $\mathbb{R}^3$  for the usual dot product. (b) Find the coordinates of  $\mathbf{v} = (1, 1, 1)^T$  relative to this basis. (c) Verify formula (4.5) in this particular case.

4.1.23. Let  $\mathbb{R}^2$  have the inner product defined by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ .

- (a) Show that  $\mathbf{v}_1 = (1, 1)^T$ ,  $\mathbf{v}_2 = (-2, 1)^T$  form an orthogonal basis. (b) Write the vector  $\mathbf{v} = (3, 2)^T$  as a linear combination of  $\mathbf{v}_1, \mathbf{v}_2$  using the orthogonality formula (4.7). (c) Verify the formula (4.8) for  $\|\mathbf{v}\|$ . (d) Find an orthonormal basis  $\mathbf{u}_1, \mathbf{u}_2$  for this inner product. (e) Write  $\mathbf{v}$  as a linear combination of the orthonormal basis, and verify (4.5).

◇ 4.1.24. (a) Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be an orthonormal basis of a finite-dimensional inner product space  $V$ .

Let  $\mathbf{v} = c_1 \mathbf{u}_1 + \dots + c_n \mathbf{u}_n$  and  $\mathbf{w} = d_1 \mathbf{u}_1 + \dots + d_n \mathbf{u}_n$  be any two elements of  $V$ .

Prove that  $\langle \mathbf{v}, \mathbf{w} \rangle = c_1 d_1 + \dots + c_n d_n$ .

- (b) Write down the corresponding inner product formula for an orthogonal basis.

4.1.25. Find an example that demonstrates why equation (4.5) is not valid for a non-orthonormal basis.

4.1.26. Use orthogonality to write the polynomials 1,  $x$  and  $x^2$  as linear combinations of the orthogonal basis (4.1).

4.1.27. (a) Prove that the polynomials  $P_0(t) = 1$ ,  $P_1(t) = t$ ,  $P_2(t) = t^2 - \frac{1}{3}$ ,  $P_3(t) = t^3 - \frac{3}{5}t$ , form an orthogonal basis for the vector space  $\mathcal{P}^{(3)}$  of cubic polynomials for the  $L^2$  inner product  $\langle f, g \rangle = \int_{-1}^1 f(t) g(t) dt$ . (b) Find an orthonormal basis of  $\mathcal{P}^{(3)}$ . (c) Write  $t^3$  as a linear combination of  $P_0, P_1, P_2, P_3$  using the orthogonal basis formula (4.7).

4.1.28. (a) Prove that the polynomials  $P_0(t) = 1$ ,  $P_1(t) = t - \frac{2}{3}$ ,  $P_2(t) = t^2 - \frac{6}{5}t + \frac{3}{10}$ , form an orthogonal basis for  $\mathcal{P}^{(2)}$  with respect to the weighted inner product

$\langle f, g \rangle = \int_0^1 f(t) g(t) t dt$ . (b) Find the corresponding orthonormal basis.

- (c) Write  $t^2$  as a linear combination of  $P_0, P_1, P_2$  using the orthogonal basis formula (4.7).

4.1.29. Write the following trigonometric polynomials in terms of the basis functions (4.11):

- (a)  $\cos^2 x$ , (b)  $\cos x \sin x$ , (c)  $\sin^3 x$ , (d)  $\cos^2 x \sin^3 x$ , (e)  $\cos^4 x$ .

*Hint:* You can use complex exponentials to simplify the inner product integrals.

4.1.30. Write down an orthonormal basis of the space of trigonometric polynomials  $\mathcal{T}^{(n)}$  with respect to the  $L^2$  inner product  $\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) g(x) dx$ .

◇ 4.1.31. Show that the  $2n+1$  complex exponentials  $e^{ikx}$  for  $k = -n, -n+1, \dots, -1, 0, 1, \dots, n$ , form an orthonormal basis for the space of complex-valued trigonometric polynomials under the Hermitian inner product  $\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} dx$ .

◇ 4.1.32. Prove the trigonometric integral identities (4.13). *Hint:* You can either use a trigonometric summation identity, or, if you can't remember the right one, use Euler's formula (3.94) to rewrite sine and cosine as combinations of complex exponentials.

◇ 4.1.33. Fill in the complete details of the proof of Theorem 4.9.

## 4.2 The Gram–Schmidt Process

Once we become convinced of the utility of orthogonal and orthonormal bases, a natural question arises: How can we construct them? A practical algorithm was first discovered by the French mathematician Pierre–Simon Laplace in the eighteenth century. Today the algorithm is known as the *Gram–Schmidt process*, after its rediscovery by Gram, whom we already met in Chapter 3, and the twentieth-century German mathematician Erhard

Schmidt. The Gram–Schmidt process is one of the premier algorithms of applied and computational linear algebra.

Let  $W$  denote a finite-dimensional inner product space. (To begin with, you might wish to think of  $W$  as a subspace of  $\mathbb{R}^m$ , equipped with the standard Euclidean dot product, although the algorithm will be formulated in complete generality.) We assume that we already know some basis  $\mathbf{w}_1, \dots, \mathbf{w}_n$  of  $W$ , where  $n = \dim W$ . Our goal is to use this information to construct an orthogonal basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ .

We will construct the orthogonal basis elements one by one. Since initially we are not worrying about normality, there are no conditions on the first orthogonal basis element  $\mathbf{v}_1$ , and so there is no harm in choosing

$$\mathbf{v}_1 = \mathbf{w}_1.$$

Note that  $\mathbf{v}_1 \neq \mathbf{0}$ , since  $\mathbf{w}_1$  appears in the original basis. Starting with  $\mathbf{w}_2$ , the second basis vector  $\mathbf{v}_2$  must be orthogonal to the first:  $\langle \mathbf{v}_2, \mathbf{v}_1 \rangle = 0$ . Let us try to arrange this by subtracting a suitable multiple of  $\mathbf{v}_1$ , and set

$$\mathbf{v}_2 = \mathbf{w}_2 - c \mathbf{v}_1,$$

where  $c$  is a scalar to be determined. The orthogonality condition

$$0 = \langle \mathbf{v}_2, \mathbf{v}_1 \rangle = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle - c \langle \mathbf{v}_1, \mathbf{v}_1 \rangle = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle - c \|\mathbf{v}_1\|^2$$

requires that  $c = \langle \mathbf{w}_2, \mathbf{v}_1 \rangle / \|\mathbf{v}_1\|^2$ , and therefore

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\langle \mathbf{w}_2, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1. \quad (4.17)$$

Linear independence of  $\mathbf{v}_1 = \mathbf{w}_1$  and  $\mathbf{w}_2$  ensures that  $\mathbf{v}_2 \neq \mathbf{0}$ . (Check!)

Next, we construct

$$\mathbf{v}_3 = \mathbf{w}_3 - c_1 \mathbf{v}_1 - c_2 \mathbf{v}_2$$

by subtracting suitable multiples of the first two orthogonal basis elements from  $\mathbf{w}_3$ . We want  $\mathbf{v}_3$  to be orthogonal to both  $\mathbf{v}_1$  and  $\mathbf{v}_2$ . Since we already arranged that  $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$ , this requires

$$0 = \langle \mathbf{v}_3, \mathbf{v}_1 \rangle = \langle \mathbf{w}_3, \mathbf{v}_1 \rangle - c_1 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle, \quad 0 = \langle \mathbf{v}_3, \mathbf{v}_2 \rangle = \langle \mathbf{w}_3, \mathbf{v}_2 \rangle - c_2 \langle \mathbf{v}_2, \mathbf{v}_2 \rangle,$$

and hence

$$c_1 = \frac{\langle \mathbf{w}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2}, \quad c_2 = \frac{\langle \mathbf{w}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2}.$$

Therefore, the next orthogonal basis vector is given by the formula

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\langle \mathbf{w}_3, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\langle \mathbf{w}_3, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} \mathbf{v}_2.$$

Since  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are linear combinations of  $\mathbf{w}_1$  and  $\mathbf{w}_2$ , we must have  $\mathbf{v}_3 \neq \mathbf{0}$ , since otherwise, this would imply that  $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3$  are linearly dependent, and hence could not come from a basis.

Continuing in the same manner, suppose we have already constructed the mutually orthogonal vectors  $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$  as linear combinations of  $\mathbf{w}_1, \dots, \mathbf{w}_{k-1}$ . The next orthogonal basis element  $\mathbf{v}_k$  will be obtained from  $\mathbf{w}_k$  by subtracting off a suitable linear combination of the previous orthogonal basis elements:

$$\mathbf{v}_k = \mathbf{w}_k - c_1 \mathbf{v}_1 - \dots - c_{k-1} \mathbf{v}_{k-1}.$$

Since  $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$  are already orthogonal, the orthogonality constraint

$$0 = \langle \mathbf{v}_k, \mathbf{v}_j \rangle = \langle \mathbf{w}_k, \mathbf{v}_j \rangle - c_j \langle \mathbf{v}_j, \mathbf{v}_j \rangle$$

requires

$$c_j = \frac{\langle \mathbf{w}_k, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \quad \text{for} \quad j = 1, \dots, k-1. \quad (4.18)$$

In this fashion, we establish the general *Gram–Schmidt formula*

$$\mathbf{v}_k = \mathbf{w}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{w}_k, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \mathbf{v}_j, \quad k = 1, \dots, n. \quad (4.19)$$

The iterative Gram–Schmidt process (4.19), where we start with  $\mathbf{v}_1 = \mathbf{w}_1$  and successively construct  $\mathbf{v}_2, \dots, \mathbf{v}_n$ , defines an explicit, recursive procedure for constructing the desired orthogonal basis vectors. If we are actually after an orthonormal basis  $\mathbf{u}_1, \dots, \mathbf{u}_n$ , we merely normalize the resulting orthogonal basis vectors, setting  $\mathbf{u}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$  for each  $k = 1, \dots, n$ .

**Example 4.13.** The vectors

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}, \quad (4.20)$$

are readily seen to form a basis<sup>†</sup> of  $\mathbb{R}^3$ . To construct an orthogonal basis (with respect to the standard dot product) using the Gram–Schmidt process, we begin by setting

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

The next basis vector is

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} - \frac{-1}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix}.$$

The last orthogonal basis vector is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} - \frac{-3}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} - \frac{7}{14} \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{3}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

The reader can easily validate the orthogonality of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ .

An orthonormal basis is obtained by dividing each vector by its length. Since

$$\|\mathbf{v}_1\| = \sqrt{3}, \quad \|\mathbf{v}_2\| = \sqrt{\frac{14}{3}}, \quad \|\mathbf{v}_3\| = \sqrt{\frac{7}{2}}.$$

---

<sup>†</sup> This will, in fact, be a consequence of the successful completion of the Gram–Schmidt process and does not need to be checked in advance. If the given vectors were not linearly independent, then eventually one of the Gram–Schmidt vectors would vanish,  $\mathbf{v}_k = \mathbf{0}$ , and the iterative algorithm would break down.

we produce the corresponding orthonormal basis vectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}. \quad (4.21)$$

**Example 4.14.** Here is a typical problem: find an orthonormal basis, with respect to the dot product, for the subspace  $W \subset \mathbb{R}^4$  consisting of all vectors that are orthogonal to the given vector  $\mathbf{a} = (1, 2, -1, -3)^T$ . The first task is to find a basis for the subspace. Now, a vector  $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$  is orthogonal to  $\mathbf{a}$  if and only if

$$\mathbf{x} \cdot \mathbf{a} = x_1 + 2x_2 - x_3 - 3x_4 = 0.$$

Solving this homogeneous linear system by the usual method, we observe that the free variables are  $x_2, x_3, x_4$ , and so a (non-orthogonal) basis for the subspace is

$$\mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

To obtain an orthogonal basis, we apply the Gram–Schmidt process. First,

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

The next element is

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \frac{-2}{5} \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix}.$$

The last element of our orthogonal basis is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{-6}{5} \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \frac{\frac{3}{5}}{\frac{6}{5}} \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ -\frac{1}{2} \\ 1 \end{pmatrix}.$$

An orthonormal basis can then be obtained by dividing each  $\mathbf{v}_i$  by its length:

$$\mathbf{u}_1 = \begin{pmatrix} -\frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{30}} \\ \frac{5}{\sqrt{30}} \\ 0 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \end{pmatrix}. \quad (4.22)$$

**Remark.** The orthonormal basis produced by the Gram–Schmidt process depends on the order of the vectors in the original basis. Different orderings produce different orthonormal bases.

The Gram–Schmidt process has one final important consequence. According to Theorem 2.29, every finite-dimensional vector space — except  $\{\mathbf{0}\}$  — admits a basis. Given an inner product, the Gram–Schmidt process enables one to construct an orthogonal and even orthonormal basis of the space. Therefore, we have, in fact, implemented a constructive proof of the existence of orthogonal and orthonormal bases of an arbitrary finite-dimensional inner product space.

**Theorem 4.15.** Every non-zero finite-dimensional inner product space has an orthonormal basis.

In fact, if its dimension is  $> 1$ , then the inner product space has infinitely many orthonormal bases.

## Exercises

*Note:* For Exercises #1–7 use the Euclidean dot product on  $\mathbb{R}^n$ .

4.2.1. Use the Gram–Schmidt process to determine an orthonormal basis for  $\mathbb{R}^3$  starting with the following sets of vectors:

$$(a) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}; \quad (b) \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}; \quad (c) \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ -1 \end{pmatrix}.$$

4.2.2. Use the Gram–Schmidt process to construct an orthonormal basis for  $\mathbb{R}^4$  starting with the following sets of vectors: (a)  $(1, 0, 1, 0)^T, (0, 1, 0, -1)^T, (1, 0, 0, 1)^T, (1, 1, 1, 1)^T$ ; (b)  $(1, 0, 0, 1)^T, (4, 1, 0, 0)^T, (1, 0, 2, 1)^T, (0, 2, 0, 1)^T$ .

4.2.3. Try the Gram–Schmidt procedure on the vectors  $\begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ -2 \\ 1 \end{pmatrix}$ .

What happens? Can you explain why you are unable to complete the algorithm?

4.2.4. Use the Gram–Schmidt process to construct an orthonormal basis for the following subspaces of  $\mathbb{R}^3$ : (a) the plane spanned by  $(0, 2, 1)^T, (1, -2, -1)^T$ ; (b) the plane defined by the equation  $2x - y + 3z = 0$ ; (c) the set of all vectors orthogonal to  $(1, -1, -2)^T$ .

4.2.5. Find an orthogonal basis of the subspace spanned by the vectors  $\mathbf{w}_1 = (1, -1, -1, 1, 1)^T$ ,  $\mathbf{w}_2 = (2, 1, 4, -4, 2)^T$ , and  $\mathbf{w}_3 = (5, -4, -3, 7, 1)^T$ .

4.2.6. Find an orthonormal basis for the following subspaces of  $\mathbb{R}^4$ : (a) the span of the vectors  $\begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 2 \\ 1 \end{pmatrix}$ ; (b) the kernel of the matrix  $\begin{pmatrix} 2 & 1 & 0 & -1 \\ 3 & 2 & -1 & -1 \end{pmatrix}$ ; (c) the coimage

of the preceding matrix; (d) the image of the matrix  $\begin{pmatrix} 1 & -2 & 2 \\ 2 & -4 & 1 \\ 0 & 0 & -1 \\ -2 & 4 & 5 \end{pmatrix}$ ; (e) the cokernel

of the preceding matrix; (f) the set of all vectors orthogonal to  $(1, 1, -1, -1)^T$ .

4.2.7. Find orthonormal bases for the four fundamental subspaces associated with the following matrices:

$$(a) \begin{pmatrix} 1 & -1 \\ -3 & 3 \end{pmatrix}, \quad (b) \begin{pmatrix} -1 & 0 & 2 \\ 1 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ -1 & 2 & 0 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 1 \\ -1 & 0 & -2 \\ 1 & -2 & 3 \end{pmatrix}.$$

4.2.8. Construct an orthonormal basis of  $\mathbb{R}^2$  for the nonstandard inner products

$$(a) \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix} \mathbf{y}, (b) \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix} \mathbf{y}, (c) \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} \mathbf{y}.$$

4.2.9. Construct an orthonormal basis for  $\mathbb{R}^3$  with respect to the inner products defined by the

following positive definite matrices: (a)  $\begin{pmatrix} 4 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 3 & -1 & 1 \\ -1 & 4 & -2 \\ 1 & -2 & 4 \end{pmatrix}$ .

4.2.10. Redo Exercise 4.2.1 using

(i) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 3v_1 w_1 + 2v_2 w_2 + v_3 w_3$ ;

(ii) the inner product induced by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ .

◇ 4.2.11. (a) How many orthonormal bases does  $\mathbb{R}$  have? (b) What about  $\mathbb{R}^2$ ? (c) Does your answer change if you use a different inner product? Justify your answers.

4.2.12. *True or false:* Reordering the original basis before starting the Gram–Schmidt process leads to the same orthogonal basis.

◇ 4.2.13. Suppose that  $W \subsetneq \mathbb{R}^n$  is a proper subspace, and  $\mathbf{u}_1, \dots, \mathbf{u}_m$  forms an orthonormal basis of  $W$ . Prove that there exist vectors  $\mathbf{u}_{m+1}, \dots, \mathbf{u}_n \in \mathbb{R}^n \setminus W$  such that the complete collection  $\mathbf{u}_1, \dots, \mathbf{u}_n$  forms an orthonormal basis for  $\mathbb{R}^n$ . *Hint:* Begin with Exercise 2.4.20.

◇ 4.2.14. Verify that the Gram–Schmidt formula (4.19) also produce an orthogonal basis of a complex vector space under a Hermitian inner product.

4.2.15. (a) Apply the complex Gram–Schmidt algorithm from Exercise 4.2.14 to produce an orthonormal basis starting with the vectors  $(1 + i, 1 - i)^T, (1 - 2i, 5i)^T \in \mathbb{C}^2$ .

(b) Do the same for  $(1 + i, 1 - i, 2 - i)^T, (1 + 2i, -2i, 2 - i)^T, (1, 1 - 2i, i)^T \in \mathbb{C}^3$ .

4.2.16. Use the complex Gram–Schmidt algorithm from Exercise 4.2.14 to construct orthonormal bases for (a) the subspace spanned by  $(1 - i, 1, 0)^T, (0, 3 - i, 2i)^T$ ; (b) the set of solutions to  $(2 - i)x - 2iy + (1 - 2i)z = 0$ ; (c) the subspace spanned by  $(-i, 1, -1, i)^T, (0, 2i, 1 - i, -1 + i)^T, (1, i, -i, 1 - 2i)^T$ .

## Modifications of the Gram–Schmidt Process

With the basic Gram–Schmidt algorithm now in hand, it is worth looking at a couple of reformulations that have both practical and theoretical advantages. The first can be used to construct the orthonormal basis vectors  $\mathbf{u}_1, \dots, \mathbf{u}_n$  directly from the basis  $\mathbf{w}_1, \dots, \mathbf{w}_n$ .

We begin by replacing each orthogonal basis vector in the basic Gram–Schmidt formula (4.19) by its normalized version  $\mathbf{u}_j = \mathbf{v}_j / \|\mathbf{v}_j\|$ . The original basis vectors can be expressed in terms of the orthonormal basis via a “triangular” system

$$\begin{aligned} \mathbf{w}_1 &= r_{11} \mathbf{u}_1, \\ \mathbf{w}_2 &= r_{12} \mathbf{u}_1 + r_{22} \mathbf{u}_2, \\ \mathbf{w}_3 &= r_{13} \mathbf{u}_1 + r_{23} \mathbf{u}_2 + r_{33} \mathbf{u}_3, \\ &\vdots \quad \vdots \quad \vdots \quad \ddots \\ \mathbf{w}_n &= r_{1n} \mathbf{u}_1 + r_{2n} \mathbf{u}_2 + \cdots + r_{nn} \mathbf{u}_n. \end{aligned} \tag{4.23}$$

The coefficients  $r_{ij}$  can, in fact, be computed directly from these formulas. Indeed, taking the inner product of the equation for  $\mathbf{w}_j$  with the orthonormal basis vector  $\mathbf{u}_i$  for  $i \leq j$ ,

we obtain, in view of the orthonormality constraints (4.6),

$$\langle \mathbf{w}_j, \mathbf{u}_i \rangle = \langle r_{1j} \mathbf{u}_1 + \cdots + r_{jj} \mathbf{u}_j, \mathbf{u}_i \rangle = r_{1j} \langle \mathbf{u}_1, \mathbf{u}_i \rangle + \cdots + r_{jj} \langle \mathbf{u}_n, \mathbf{u}_i \rangle = r_{ij},$$

and hence

$$r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle. \quad (4.24)$$

On the other hand, according to (4.5),

$$\|\mathbf{w}_j\|^2 = \|r_{1j} \mathbf{u}_1 + \cdots + r_{jj} \mathbf{u}_j\|^2 = r_{1j}^2 + \cdots + r_{j-1,j}^2 + r_{jj}^2. \quad (4.25)$$

The pair of equations (4.24–25) can be rearranged to devise a recursive procedure to compute the orthonormal basis. We begin by setting  $r_{11} = \|\mathbf{w}_1\|$  and so  $\mathbf{u}_1 = \mathbf{w}_1/r_{11}$ . At each subsequent stage  $j \geq 2$ , we assume that we have already constructed  $\mathbf{u}_1, \dots, \mathbf{u}_{j-1}$ . We then compute

$$r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle, \quad \text{for each } i = 1, \dots, j-1. \quad (4.26)$$

We obtain the next orthonormal basis vector  $\mathbf{u}_j$  by computing

$$r_{jj} = \sqrt{\|\mathbf{w}_j\|^2 - r_{1j}^2 - \cdots - r_{j-1,j}^2}, \quad \mathbf{u}_j = \frac{\mathbf{w}_j - r_{1j} \mathbf{u}_1 - \cdots - r_{j-1,j} \mathbf{u}_{j-1}}{r_{jj}}. \quad (4.27)$$

Running through the formulas (4.26–27) for  $j = 1, \dots, n$  leads to the *same* orthonormal basis  $\mathbf{u}_1, \dots, \mathbf{u}_n$  produced by the previous version of the Gram–Schmidt procedure.

**Example 4.16.** Let us apply the revised algorithm to the vectors

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix},$$

of Example 4.13. To begin, we set

$$r_{11} = \|\mathbf{w}_1\| = \sqrt{3}, \quad \mathbf{u}_1 = \frac{\mathbf{w}_1}{r_{11}} = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}.$$

The next step is to compute

$$r_{12} = \langle \mathbf{w}_2, \mathbf{u}_1 \rangle = -\frac{1}{\sqrt{3}}, \quad r_{22} = \sqrt{\|\mathbf{w}_2\|^2 - r_{12}^2} = \sqrt{\frac{14}{3}}, \quad \mathbf{u}_2 = \frac{\mathbf{w}_2 - r_{12} \mathbf{u}_1}{r_{22}} = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}.$$

The final step yields

$$r_{13} = \langle \mathbf{w}_3, \mathbf{u}_1 \rangle = -\sqrt{3}, \quad r_{23} = \langle \mathbf{w}_3, \mathbf{u}_2 \rangle = \sqrt{\frac{21}{2}},$$

$$r_{33} = \sqrt{\|\mathbf{w}_3\|^2 - r_{13}^2 - r_{23}^2} = \sqrt{\frac{7}{2}}, \quad \mathbf{u}_3 = \frac{\mathbf{w}_3 - r_{13} \mathbf{u}_1 - r_{23} \mathbf{u}_2}{r_{33}} = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}.$$

As advertised, the result is the same orthonormal basis vectors that we previously found in Example 4.13.

For hand computations, the original version (4.19) of the Gram–Schmidt process is slightly easier — even if one does ultimately want an orthonormal basis — since it avoids

the square roots that are ubiquitous in the orthonormal version (4.26–27). On the other hand, for numerical implementation on a computer, the orthonormal version is a bit faster, since it involves fewer arithmetic operations.

However, in practical, large-scale computations, both versions of the Gram–Schmidt process suffer from a serious flaw. They are subject to numerical instabilities, and so accumulating round-off errors may seriously corrupt the computations, leading to inaccurate, non-orthogonal vectors. Fortunately, there is a simple rearrangement of the calculation that ameliorates this difficulty and leads to the numerically robust algorithm that is most often used in practice, [21, 40, 66]. The idea is to treat the vectors simultaneously rather than sequentially, making full use of the orthonormal basis vectors as they arise. More specifically, the algorithm begins as before — we take  $\mathbf{u}_1 = \mathbf{w}_1 / \|\mathbf{w}_1\|$ . We then subtract off the appropriate multiples of  $\mathbf{u}_1$  from *all* of the remaining basis vectors so as to arrange their orthogonality to  $\mathbf{u}_1$ . This is accomplished by setting

$$\mathbf{w}_k^{(2)} = \mathbf{w}_k - \langle \mathbf{w}_k, \mathbf{u}_1 \rangle \mathbf{u}_1 \quad \text{for } k = 2, \dots, n.$$

The second orthonormal basis vector  $\mathbf{u}_2 = \mathbf{w}_2^{(2)} / \|\mathbf{w}_2^{(2)}\|$  is then obtained by normalizing. We next modify the remaining  $\mathbf{w}_3^{(2)}, \dots, \mathbf{w}_n^{(2)}$  to produce vectors

$$\mathbf{w}_k^{(3)} = \mathbf{w}_k^{(2)} - \langle \mathbf{w}_k^{(2)}, \mathbf{u}_2 \rangle \mathbf{u}_2, \quad k = 3, \dots, n,$$

that are orthogonal to both  $\mathbf{u}_1$  and  $\mathbf{u}_2$ . Then  $\mathbf{u}_3 = \mathbf{w}_3^{(3)} / \|\mathbf{w}_3^{(3)}\|$  is the next orthonormal basis element, and the process continues. The full algorithm starts with the initial basis vectors  $\mathbf{w}_j = \mathbf{w}_j^{(1)}$ ,  $j = 1, \dots, n$ , and then recursively computes

$$\mathbf{u}_j = \frac{\mathbf{w}_j^{(j)}}{\|\mathbf{w}_j^{(j)}\|}, \quad \mathbf{w}_k^{(j+1)} = \mathbf{w}_k^{(j)} - \langle \mathbf{w}_k^{(j)}, \mathbf{u}_j \rangle \mathbf{u}_j, \quad j = 1, \dots, n, \quad k = j+1, \dots, n. \quad (4.28)$$

(In the final phase, when  $j = n$ , the second formula is no longer needed.) The result is a numerically stable computation of the *same* orthonormal basis vectors  $\mathbf{u}_1, \dots, \mathbf{u}_n$ .

**Example 4.17.** Let us apply the stable Gram–Schmidt process to the basis vectors

$$\mathbf{w}_1^{(1)} = \mathbf{w}_1 = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{w}_2^{(1)} = \mathbf{w}_2 = \begin{pmatrix} 0 \\ 4 \\ -1 \end{pmatrix}, \quad \mathbf{w}_3^{(1)} = \mathbf{w}_3 = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}.$$

The first orthonormal basis vector is  $\mathbf{u}_1 = \frac{\mathbf{w}_1^{(1)}}{\|\mathbf{w}_1^{(1)}\|} = \begin{pmatrix} \frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$ . Next, we compute

$$\mathbf{w}_2^{(2)} = \mathbf{w}_2^{(1)} - \langle \mathbf{w}_2^{(1)}, \mathbf{u}_1 \rangle \mathbf{u}_1 = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}, \quad \mathbf{w}_3^{(2)} = \mathbf{w}_3^{(1)} - \langle \mathbf{w}_3^{(1)}, \mathbf{u}_1 \rangle \mathbf{u}_1 = \begin{pmatrix} -1 \\ 0 \\ -2 \end{pmatrix}.$$

The second orthonormal basis vector is  $\mathbf{u}_2 = \frac{\mathbf{w}_2^{(2)}}{\|\mathbf{w}_2^{(2)}\|} = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$ . Finally,

$$\mathbf{w}_3^{(3)} = \mathbf{w}_3^{(2)} - \langle \mathbf{w}_3^{(2)}, \mathbf{u}_2 \rangle \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ -2 \end{pmatrix}, \quad \mathbf{u}_3 = \frac{\mathbf{w}_3^{(3)}}{\|\mathbf{w}_3^{(3)}\|} = \begin{pmatrix} -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \\ -\frac{2\sqrt{2}}{3} \end{pmatrix}.$$

The resulting vectors  $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$  form the desired orthonormal basis.

## Exercises

- 4.2.17. Use the modified Gram-Schmidt process (4.26–27) to produce orthonormal bases for the spaces spanned by the following vectors: (a)  $\begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}$ , (b)  $\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 2 \\ 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix}$ , (e)  $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix}$ .
- 4.2.18. Repeat Exercise 4.2.17 using the numerically stable algorithm (4.28) and check that you get the same result. Which of the two algorithms was easier for you to implement?
- 4.2.19. Redo each of the exercises in the preceding subsection by implementing the numerically stable Gram-Schmidt process (4.28) instead, and verify that you end up with the same orthonormal basis.
- ◊ 4.2.20. Prove that (4.28) does indeed produce an orthonormal basis. Explain why the result is the same orthonormal basis as the ordinary Gram-Schmidt method.
- 4.2.21. Let  $\mathbf{w}_j^{(j)}$  be the vectors in the stable Gram-Schmidt algorithm (4.28). Prove that the coefficients in (4.23) are given by  $r_{ii} = \|\mathbf{w}_i^{(i)}\|$ , and  $r_{ij} = \langle \mathbf{w}_j^{(i)}, \mathbf{u}_i \rangle$  for  $i < j$ .

## 4.3 Orthogonal Matrices

Matrices whose columns form an orthonormal basis of  $\mathbb{R}^n$  relative to the standard Euclidean dot product play a distinguished role. Such “orthogonal matrices” appear in a wide range of applications in geometry, physics, quantum mechanics, crystallography, partial differential equations, [61], symmetry theory, [60], and special functions, [59]. Rotational motions of bodies in three-dimensional space are described by orthogonal matrices, and hence they lie at the foundations of rigid body mechanics, [31], including satellites, airplanes, drones, and underwater vehicles, as well as three-dimensional computer graphics and animation for video games and movies, [5]. Furthermore, orthogonal matrices are an essential ingredient in one of the most important methods of numerical linear algebra: the  $QR$  algorithm for computing eigenvalues of matrices, to be presented in Section 9.5.

**Definition 4.18.** A square matrix  $Q$  is called *orthogonal* if it satisfies

$$Q^T Q = Q Q^T = I. \quad (4.29)$$

The orthogonality condition implies that one can easily invert an orthogonal matrix:

$$Q^{-1} = Q^T. \quad (4.30)$$

In fact, the two conditions are equivalent, and hence a matrix is orthogonal if and only if its inverse is equal to its transpose. In particular, the identity matrix  $I$  is orthogonal. Also note that if  $Q$  is orthogonal, so is  $Q^T$ . The second important characterization of orthogonal matrices relates them directly to orthonormal bases.

**Proposition 4.19.** A matrix  $Q$  is orthogonal if and only if its columns form an orthonormal basis with respect to the Euclidean dot product on  $\mathbb{R}^n$ .

*Proof:* Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be the columns of  $Q$ . Then  $\mathbf{u}_1^T, \dots, \mathbf{u}_n^T$  are the rows of the transposed matrix  $Q^T$ . The  $(i, j)$  entry of the product  $Q^T Q$  is given as the product of the  $i^{\text{th}}$  row of  $Q^T$  and the  $j^{\text{th}}$  column of  $Q$ . Thus, the orthogonality requirement (4.29) implies  $\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$  which are precisely the conditions (4.6) for  $\mathbf{u}_1, \dots, \mathbf{u}_n$  to form an orthonormal basis.  $Q.E.D.$

In particular, the columns of the identity matrix produce the standard basis  $\mathbf{e}_1, \dots, \mathbf{e}_n$  of  $\mathbb{R}^n$ . Also, the rows of an orthogonal matrix  $Q$  also produce an (in general different) orthonormal basis.

**Warning.** Technically, we should be referring to an “orthonormal” matrix, not an “orthogonal” matrix. But the terminology is so standard throughout mathematics and physics that we have no choice but to adopt it here. There is no commonly accepted name for a matrix whose columns form an orthogonal but not orthonormal basis.

**Example 4.20.** A  $2 \times 2$  matrix  $Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is orthogonal if and only if its columns  $\mathbf{u}_1 = \begin{pmatrix} a \\ c \end{pmatrix}$ ,  $\mathbf{u}_2 = \begin{pmatrix} b \\ d \end{pmatrix}$ , form an orthonormal basis of  $\mathbb{R}^2$ . Equivalently, the requirement  $Q^T Q = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ ,

implies that its entries must satisfy the algebraic equations

$$a^2 + c^2 = 1, \quad ab + cd = 0, \quad b^2 + d^2 = 1.$$

The first and last equations say that the points  $(a, c)^T$  and  $(b, d)^T$  lie on the unit circle in  $\mathbb{R}^2$ , and so

$$a = \cos \theta, \quad c = \sin \theta, \quad b = \cos \psi, \quad d = \sin \psi,$$

for some choice of angles  $\theta, \psi$ . The remaining orthogonality condition is

$$0 = ab + cd = \cos \theta \cos \psi + \sin \theta \sin \psi = \cos(\theta - \psi),$$

which implies that  $\theta$  and  $\psi$  differ by a right angle:  $\psi = \theta \pm \frac{1}{2}\pi$ . The  $\pm$  sign leads to two cases:

$$b = -\sin \theta, \quad d = \cos \theta, \quad \text{or} \quad b = \sin \theta, \quad d = -\cos \theta.$$

As a result, every  $2 \times 2$  orthogonal matrix has one of two possible forms

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}, \quad \text{where } 0 \leq \theta < 2\pi. \quad (4.31)$$

The corresponding orthonormal bases are illustrated in Figure 4.2. The former is a right-handed basis, as defined in Exercise 2.4.7, and can be obtained from the standard basis  $\mathbf{e}_1, \mathbf{e}_2$  by a rotation through angle  $\theta$ , while the latter has the opposite, reflected orientation.

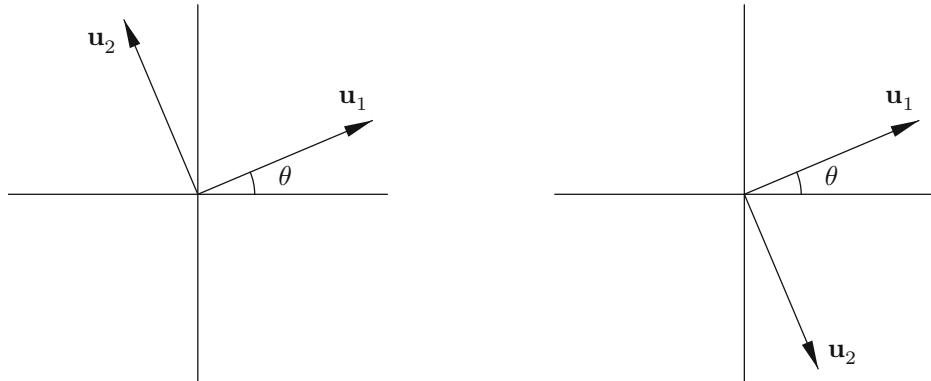


Figure 4.2. Orthonormal Bases in  $\mathbb{R}^2$ .

**Example 4.21.** A  $3 \times 3$  orthogonal matrix  $Q = (\mathbf{u}_1 \ \mathbf{u}_2 \ \mathbf{u}_3)$  is prescribed by 3 mutually perpendicular vectors of unit length in  $\mathbb{R}^3$ . For instance, the orthonormal basis constructed

in (4.21) corresponds to the orthogonal matrix  $Q = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix}$ . A complete

list of  $3 \times 3$  orthogonal matrices can be found in Exercises 4.3.4 and 4.3.5.

**Lemma 4.22.** An orthogonal matrix  $Q$  has determinant  $\det Q = \pm 1$ .

*Proof:* Taking the determinant of (4.29), and using the determinantal formulas (1.85), (1.89), shows that

$$1 = \det \mathbf{I} = \det(Q^T Q) = \det Q^T \det Q = (\det Q)^2,$$

which immediately proves the lemma. *Q.E.D.*

An orthogonal matrix is called *proper* or *special* if it has determinant  $+1$ . Geometrically, the columns of a proper orthogonal matrix form a right-handed basis of  $\mathbb{R}^n$ , as defined in Exercise 2.4.7. An *improper* orthogonal matrix, with determinant  $-1$ , corresponds to a left handed basis that lives in a mirror-image world.

**Proposition 4.23.** The product of two orthogonal matrices is also orthogonal.

*Proof:* If

$$Q_1^T Q_1 = \mathbf{I} = Q_2^T Q_2, \quad \text{then} \quad (Q_1 Q_2)^T (Q_1 Q_2) = Q_2^T Q_1^T Q_1 Q_2 = Q_2^T Q_2 = \mathbf{I},$$

and so the product matrix  $Q_1 Q_2$  is also orthogonal. *Q.E.D.*

This multiplicative property combined with the fact that the inverse of an orthogonal matrix is also orthogonal says that the set of all orthogonal matrices forms a *group*<sup>†</sup>. The

<sup>†</sup> The precise mathematical definition of a group can be found in Exercise 4.3.24. Although they will not play a significant role in this text, groups underlie the mathematical formalization of symmetry and, as such, form one of the most fundamental concepts in advanced mathematics and its applications, particularly quantum mechanics and modern theoretical physics, [54]. Indeed, according to the mathematician Felix Klein, cf. [92], all geometry is based on group theory.

*orthogonal group* lies at the foundation of everyday Euclidean geometry, as well as rigid body mechanics, atomic structure and chemistry, computer graphics and animation, and many other areas.

## Exercises

4.3.1. Determine which of the following matrices are (i) orthogonal; (ii) proper orthogonal.

$$(a) \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad (b) \begin{pmatrix} \frac{12}{13} & \frac{5}{13} \\ -\frac{5}{13} & \frac{12}{13} \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad (d) \begin{pmatrix} -\frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & -\frac{1}{3} \end{pmatrix},$$

$$(e) \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \end{pmatrix}, \quad (f) \begin{pmatrix} \frac{3}{5} & 0 & \frac{4}{5} \\ -\frac{4}{13} & \frac{12}{13} & \frac{3}{13} \\ -\frac{48}{65} & -\frac{5}{13} & \frac{36}{65} \end{pmatrix}, \quad (g) \begin{pmatrix} \frac{2}{3} & -\frac{\sqrt{2}}{6} & \frac{\sqrt{2}}{2} \\ -\frac{2}{3} & \frac{\sqrt{2}}{6} & \frac{\sqrt{2}}{2} \\ \frac{1}{3} & \frac{2\sqrt{2}}{3} & 0 \end{pmatrix}.$$

4.3.2. (a) Show that  $R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ , a reflection matrix, and  $Q = \begin{pmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$ ,

representing a rotation by the angle  $\theta$  around the  $z$ -axis, are both orthogonal. (b) Verify that the products  $RQ$  and  $QR$  are also orthogonal. (c) Which of the preceding matrices,  $R, Q, RQ, QR$ , are proper orthogonal?

4.3.3. *True or false:* (a) If  $Q$  is an improper  $2 \times 2$  orthogonal matrix, then  $Q^2 = I$ .

(b) If  $Q$  is an improper  $3 \times 3$  orthogonal matrix, then  $Q^2 = I$ .

◇ 4.3.4. (a) Prove that, for all  $\theta, \varphi, \psi$ ,

$$Q = \begin{pmatrix} \cos \varphi \cos \psi - \cos \theta \sin \varphi \sin \psi & \sin \varphi \cos \psi + \cos \theta \cos \varphi \sin \psi & \sin \theta \sin \psi \\ -\cos \varphi \sin \psi - \cos \theta \sin \varphi \cos \psi & -\sin \varphi \sin \psi + \cos \theta \cos \varphi \cos \psi & \sin \theta \cos \psi \\ \sin \theta \sin \varphi & -\sin \theta \cos \varphi & \cos \theta \end{pmatrix}$$

is a proper orthogonal matrix. (b) Write down a formula for  $Q^{-1}$ .

**Remark.** It can be shown that every proper orthogonal matrix can be parameterized in this manner;  $\theta, \varphi, \psi$  are known as the *Euler angles*, and play an important role in applications in mechanics and geometry, [31; p. 147].

◇ 4.3.5. (a) Show that if  $y_1^2 + y_2^2 + y_3^2 + y_4^2 = 1$ , then the matrix

$$Q = \begin{pmatrix} y_1^2 + y_2^2 - y_3^2 - y_4^2 & 2(y_2 y_3 + y_1 y_4) & 2(y_2 y_4 - y_1 y_3) \\ 2(y_2 y_3 - y_1 y_4) & y_1^2 - y_2^2 + y_3^2 - y_4^2 & 2(y_3 y_4 + y_1 y_2) \\ 2(y_2 y_4 + y_1 y_3) & 2(y_3 y_4 - y_1 y_2) & y_1^2 - y_2^2 - y_3^2 + y_4^2 \end{pmatrix}$$

is a proper orthogonal matrix. The numbers  $y_1, y_2, y_3, y_4$  are known as *Cayley–Klein parameters*. (b) Write down a formula for  $Q^{-1}$ . (c) Prove the formulas

$$y_1 = \cos \frac{\varphi + \psi}{2} \cos \frac{\theta}{2}, \quad y_2 = \cos \frac{\varphi - \psi}{2} \sin \frac{\theta}{2}, \quad y_3 = \sin \frac{\varphi - \psi}{2} \sin \frac{\theta}{2}, \quad y_4 = \sin \frac{\varphi + \psi}{2} \cos \frac{\theta}{2},$$

relating the Cayley–Klein parameters and the Euler angles of Exercise 4.3.4, cf. [31; §§4–5].

◇ 4.3.6. (a) Prove that the transpose of an orthogonal matrix is also orthogonal. (b) Explain why the rows of an  $n \times n$  orthogonal matrix also form an orthonormal basis of  $\mathbb{R}^n$ .

4.3.7. Prove that the inverse of an orthogonal matrix is orthogonal.

4.3.8. Show that if  $Q$  is a proper orthogonal matrix, and  $R$  is obtained from  $Q$  by interchanging two rows, then  $R$  is an improper orthogonal matrix.

4.3.9. Show that the product of two proper orthogonal matrices is also proper orthogonal.

What can you say about the product of two improper orthogonal matrices? What about an improper times a proper orthogonal matrix?

4.3.10. *True or false:* (a) A matrix whose columns form an orthogonal basis of  $\mathbb{R}^n$  is an orthogonal matrix. (b) A matrix whose rows form an orthonormal basis of  $\mathbb{R}^n$  is an orthogonal matrix. (c) An orthogonal matrix is symmetric if and only if it is a diagonal matrix.

4.3.11. Write down all diagonal  $n \times n$  orthogonal matrices.

◇ 4.3.12. Prove that an upper triangular matrix  $U$  is orthogonal if and only if  $U$  is a diagonal matrix. What are its diagonal entries?

4.3.13. (a) Show that the elementary row operation matrix corresponding to the interchange of two rows is an improper orthogonal matrix. (b) Are there any other orthogonal elementary matrices?

4.3.14. *True or false:* Applying an elementary row operation to an orthogonal matrix produces an orthogonal matrix.

4.3.15. (a) Prove that every permutation matrix is orthogonal. (b) How many permutation matrices of a given size are proper orthogonal?

◇ 4.3.16. (a) Prove that if  $Q$  is an orthogonal matrix, then  $\|Q\mathbf{x}\| = \|\mathbf{x}\|$  for every vector  $\mathbf{x} \in \mathbb{R}^n$ , where  $\|\cdot\|$  denotes the standard Euclidean norm. (b) Prove the converse: if  $\|Q\mathbf{x}\| = \|\mathbf{x}\|$  for all  $\mathbf{x} \in \mathbb{R}^n$ , then  $Q$  is an orthogonal matrix.

◇ 4.3.17. Show that if  $A^T = -A$  is any skew-symmetric matrix, then its *Cayley Transform*  $Q = (I - A)^{-1}(I + A)$  is an orthogonal matrix. Can you prove that  $I - A$  is always invertible?

4.3.18. Suppose  $S$  is an  $n \times n$  matrix whose columns form an orthogonal, but not orthonormal, basis of  $\mathbb{R}^n$ . (a) Find a formula for  $S^{-1}$  mimicking the formula  $Q^{-1} = Q^T$  for an orthogonal matrix. (b) Use your formula to determine the inverse of the wavelet matrix  $W$  whose columns form the orthogonal wavelet basis (4.9) of  $\mathbb{R}^4$ .

◇ 4.3.19. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  and  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be two sets of linearly independent vectors in  $\mathbb{R}^n$ . Show that all their dot products are the same, so  $\mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{w}_i \cdot \mathbf{w}_j$  for all  $i, j = 1, \dots, n$ , if and only if there is an orthogonal matrix  $Q$  such that  $\mathbf{w}_i = Q\mathbf{v}_i$  for all  $i = 1, \dots, n$ .

4.3.20. Suppose  $\mathbf{u}_1, \dots, \mathbf{u}_k$  form an orthonormal set of vectors in  $\mathbb{R}^n$  with  $k < n$ . Let  $Q = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_k)$  denote the  $n \times k$  matrix whose columns are the orthonormal vectors. (a) Prove that  $Q^T Q = I_k$ . (b) Is  $Q Q^T = I_n$ ?

◇ 4.3.21. Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  and  $\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_n$  be orthonormal bases of an inner product space  $V$ .

Prove that  $\hat{\mathbf{u}}_i = \sum_{j=1}^n q_{ij} \mathbf{u}_j$  for  $i = 1, \dots, n$ , where  $Q = (q_{ij})$  is an orthogonal matrix.

4.3.22. Let  $A$  be an  $m \times n$  matrix whose columns are nonzero, mutually orthogonal vectors in  $\mathbb{R}^m$ . (a) Explain why  $m \geq n$ . (b) Prove that  $A^T A$  is a diagonal matrix. What are the diagonal entries? (c) Is  $A A^T$  diagonal?

◇ 4.3.23. Let  $K > 0$  be a positive definite  $n \times n$  matrix. Prove that an  $n \times n$  matrix  $S$  satisfies  $S^T K S = I$  if and only if the columns of  $S$  form an orthonormal basis of  $\mathbb{R}^n$  with respect to the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T K \mathbf{w}$ .

◇ 4.3.24. *Groups:* A set of  $n \times n$  matrices  $G \subset \mathcal{M}_{n \times n}$  is said to form a *group* if

- (1) whenever  $A, B \in G$ , so is the product  $AB \in G$ , and
- (2) whenever  $A \in G$ , then  $A$  is nonsingular, and  $A^{-1} \in G$ .

- (a) Show that  $I \in G$ . (b) Prove that the following sets of  $n \times n$  matrices form a group:  
 (i) all nonsingular matrices; (ii) all nonsingular upper triangular matrices; (iii) all matrices of determinant 1; (iv) all orthogonal matrices; (v) all proper orthogonal matrices; (vi) all permutation matrices; (vii) all  $2 \times 2$  matrices with integer entries and determinant equal to 1. (c) Explain why the set of all nonsingular  $2 \times 2$  matrices with integer entries does not form a group. (d) Does the set of positive definite matrices form a group?

◇ 4.3.25. *Unitary matrices:* A complex, square matrix  $U$  is called *unitary* if it satisfies  $U^\dagger U = I$ , where  $U^\dagger = \overline{U^T}$  denotes the *Hermitian adjoint* in which one first transposes and then takes complex conjugates of all entries. (a) Show that  $U$  is a unitary matrix if and only if  $U^{-1} = U^\dagger$ . (b) Show that the following matrices are unitary and compute their inverses:

$$(i) \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \quad (ii) \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} + \frac{i}{2} & -\frac{1}{2\sqrt{3}} - \frac{i}{2} \\ \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} - \frac{i}{2} & -\frac{1}{2\sqrt{3}} + \frac{i}{2} \end{pmatrix}, \quad (iii) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{i}{2} & -\frac{1}{2} & -\frac{i}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{i}{2} & -\frac{1}{2} & \frac{i}{2} \end{pmatrix}.$$

- (c) Are the following matrices unitary?

$$(i) \begin{pmatrix} 2 & 1+2i \\ 1-2i & 3 \end{pmatrix}, \quad (ii) \frac{1}{5} \begin{pmatrix} -1+2i & -4-2i \\ 2-4i & -2-i \end{pmatrix}, \quad (iii) \begin{pmatrix} \frac{12}{13} & \frac{5}{13} \\ \frac{5}{13} & -\frac{12}{13} \end{pmatrix}.$$

- (d) Show that  $U$  is a unitary matrix if and only if its columns form an orthonormal basis of  $\mathbb{C}^n$  with respect to the Hermitian dot product. (e) Prove that the set of unitary matrices forms a group, as defined in Exercise 4.3.24.

## The QR Factorization

The Gram–Schmidt procedure for orthonormalizing bases of  $\mathbb{R}^n$  can be reinterpreted as a matrix factorization. This is more subtle than the *LU* factorization that resulted from Gaussian Elimination, but is of comparable significance, and is used in a broad range of applications in mathematics, statistics, physics, engineering, and numerical analysis.

Let  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be a basis of  $\mathbb{R}^n$ , and let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be the corresponding orthonormal basis that results from any one of the three implementations of the Gram–Schmidt process. We assemble both sets of column vectors to form nonsingular  $n \times n$  matrices

$$A = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n), \quad Q = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n).$$

Since the  $\mathbf{u}_i$  form an orthonormal basis,  $Q$  is an orthogonal matrix. In view of the matrix multiplication formula (2.13), the Gram–Schmidt equations (4.23) can be recast into an equivalent matrix form:

$$A = QR, \quad \text{where} \quad R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix} \quad (4.32)$$

is an upper triangular matrix whose entries are the coefficients in (4.26–27). Since the Gram–Schmidt process works on any basis, the only requirement on the matrix  $A$  is that its columns form a basis of  $\mathbb{R}^n$ , and hence  $A$  can be any nonsingular matrix. We have therefore established the celebrated *QR factorization* of nonsingular matrices.

**Theorem 4.24.** Every nonsingular matrix can be factored,  $A = QR$ , into the product of an orthogonal matrix  $Q$  and an upper triangular matrix  $R$ . The factorization is unique if  $R$  is *positive upper triangular*, meaning that all its diagonal entries are positive.

---

 QR Factorization of a Matrix A
 

---

```

start
  for  $j = 1$  to  $n$ 
    set  $r_{jj} = \sqrt{a_{1j}^2 + \dots + a_{nj}^2}$ 
    if  $r_{jj} = 0$ , stop; print “ $A$  has linearly dependent columns”
    else for  $i = 1$  to  $n$ 
      set  $a_{ij} = a_{ij}/r_{jj}$ 
      next  $i$ 
    for  $k = j + 1$  to  $n$ 
      set  $r_{jk} = a_{1j}a_{1k} + \dots + a_{nj}a_{nk}$ 
      for  $i = 1$  to  $n$ 
        set  $a_{ik} = a_{ik} - a_{ij}r_{jk}$ 
        next  $i$ 
      next  $k$ 
    next  $j$ 
  end

```

---

The proof of uniqueness is relegated to Exercise 4.3.30.

**Example 4.25.** The columns of the matrix  $A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \\ -1 & 2 & 3 \end{pmatrix}$  are the same basis

vectors considered in Example 4.16. The orthonormal basis (4.21) constructed using the Gram–Schmidt algorithm leads to the orthogonal and upper triangular matrices

$$Q = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix}, \quad R = \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}. \quad (4.33)$$

The reader may wish to verify that, indeed,  $A = QR$ .

While any of the three implementations of the Gram–Schmidt algorithm will produce the  $QR$  factorization of a given matrix  $A = (\mathbf{w}_1 \mathbf{w}_2 \dots \mathbf{w}_n)$ , the stable version, as encoded in equations (4.28), is the one to use in practical computations, since it is the least likely to fail due to numerical artifacts produced by round-off errors. The accompanying pseudocode program reformulates the algorithm purely in terms of the matrix entries  $a_{ij}$  of  $A$ . During the course of the algorithm, the entries of the matrix  $A$  are successively overwritten; the final result is the orthogonal matrix  $Q$  appearing in place of  $A$ . The entries  $r_{ij}$  of  $R$  must be stored separately.

**Example 4.26.** Let us factor the matrix  $A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$  using the numerically

stable  $QR$  algorithm. As in the program, we work directly on the matrix  $A$ , gradually

changing it into orthogonal form. In the first loop, we set  $r_{11} = \sqrt{5}$  to be the norm of the first column vector of  $A$ . We then normalize the first column by dividing by

$r_{11}$ ; the resulting matrix is  $\begin{pmatrix} \frac{2}{\sqrt{5}} & 1 & 0 & 0 \\ \frac{1}{\sqrt{5}} & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$ . The next entries  $r_{12} = \frac{4}{\sqrt{5}}$ ,  $r_{13} = \frac{1}{\sqrt{5}}$ ,

$r_{14} = 0$ , are obtained by taking the dot products of the first column with the other three columns. For  $j = 1, 2, 3$ , we subtract  $r_{1j}$  times the first column from the  $j^{\text{th}}$  column;

the result  $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{5} & -\frac{2}{5} & 0 \\ \frac{1}{\sqrt{5}} & \frac{6}{5} & \frac{4}{5} & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$  is a matrix whose first column is normalized to have

unit length, and whose second, third and fourth columns are orthogonal to it. In the next loop, we normalize the second column by dividing by its norm  $r_{22} = \sqrt{\frac{14}{5}}$ , and so

obtain the matrix  $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & -\frac{2}{5} & 0 \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & \frac{4}{5} & 0 \\ 0 & \frac{5}{\sqrt{70}} & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$ . We then take dot products of the second

column with the remaining two columns to produce  $r_{23} = \frac{16}{\sqrt{70}}$ ,  $r_{24} = \frac{5}{\sqrt{70}}$ . Subtracting these multiples of the second column from the third and fourth columns, we obtain

$\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{7} & \frac{3}{14} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{7} & -\frac{3}{7} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{7} & \frac{9}{14} \\ 0 & 0 & 1 & 2 \end{pmatrix}$ , which now has its first two columns orthonormalized, and or-

thogonal to the last two columns. We then normalize the third column by dividing by

$r_{33} = \sqrt{\frac{15}{7}}$ , yielding  $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{\sqrt{105}} & \frac{3}{14} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{\sqrt{105}} & -\frac{3}{7} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{\sqrt{105}} & \frac{9}{14} \\ 0 & 0 & \frac{7}{\sqrt{105}} & 2 \end{pmatrix}$ . Finally, we subtract  $r_{34} = \frac{20}{\sqrt{105}}$  times

the third column from the fourth column. Dividing the resulting fourth column by its norm  $r_{44} = \sqrt{\frac{5}{6}}$  results in the final formulas,

$$Q = \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{\sqrt{105}} & -\frac{1}{\sqrt{30}} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{\sqrt{105}} & \frac{2}{\sqrt{30}} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{\sqrt{105}} & -\frac{3}{\sqrt{30}} \\ 0 & 0 & \frac{7}{\sqrt{105}} & \frac{4}{\sqrt{30}} \end{pmatrix}, \quad R = \begin{pmatrix} \sqrt{5} & \frac{4}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ 0 & \frac{\sqrt{14}}{\sqrt{5}} & \frac{16}{\sqrt{70}} & \frac{5}{\sqrt{70}} \\ 0 & 0 & \frac{\sqrt{15}}{\sqrt{7}} & \frac{20}{\sqrt{105}} \\ 0 & 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} \end{pmatrix},$$

for the  $A = QR$  factorization.

## III-Conditioned Systems and Householder's Method

The  $QR$  factorization can be employed as an alternative to Gaussian Elimination to solve linear systems. Indeed, the system

$$A\mathbf{x} = \mathbf{b} \quad \text{becomes} \quad QR\mathbf{x} = \mathbf{b}, \quad \text{and hence} \quad R\mathbf{x} = Q^T\mathbf{b}, \quad (4.34)$$

because  $Q^{-1} = Q^T$  is an orthogonal matrix. Since  $R$  is upper triangular, the latter system can be solved for  $\mathbf{x}$  by Back Substitution. The resulting algorithm, while more expensive to compute, offers some numerical advantages over traditional Gaussian Elimination, since it is less prone to inaccuracies resulting from ill-conditioning.

**Example 4.27.** Let us apply the  $A = QR$  factorization

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \\ -1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}$$

that we found in Example 4.25 to solve the linear system  $A\mathbf{x} = (0, -4, 5)^T$ . We first compute

$$Q^T\mathbf{b} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{4}{\sqrt{42}} & \frac{1}{\sqrt{42}} & \frac{5}{\sqrt{42}} \\ \frac{2}{\sqrt{14}} & -\frac{3}{\sqrt{14}} & -\frac{1}{\sqrt{14}} \end{pmatrix} \begin{pmatrix} 0 \\ -4 \\ 5 \end{pmatrix} = \begin{pmatrix} -3\sqrt{3} \\ \frac{\sqrt{21}}{\sqrt{2}} \\ \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}.$$

We then solve the upper triangular system

$$R\mathbf{x} = \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -3\sqrt{3} \\ \frac{\sqrt{21}}{\sqrt{2}} \\ \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}$$

by Back Substitution, leading to the solution  $\mathbf{x} = (-2, 0, 1)^T$ .

In computing the  $QR$  factorization of a mildly ill-conditioned matrix, one should employ the stable version (4.28) of the Gram–Schmidt process. However, yet more recalcitrant matrices require a completely different approach to the factorization, as formulated by the mid-twentieth-century American mathematician Alston Householder. His idea was to use a sequence of certain simple orthogonal matrices to gradually convert the matrix into upper triangular form.

Consider the *Householder* or *elementary reflection matrix*

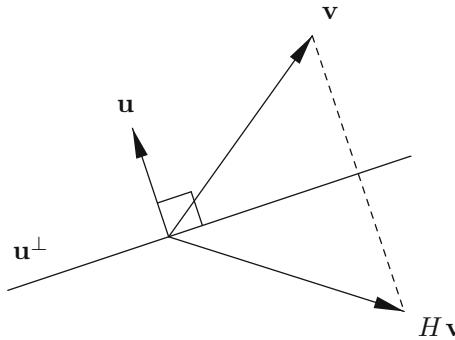
$$H = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T, \quad (4.35)$$

in which  $\mathbf{u}$  is a unit vector (in the Euclidean norm). Geometrically, the matrix  $H$  represents a reflection of vectors through the subspace

$$\mathbf{u}^\perp = \{ \mathbf{v} \mid \mathbf{v} \cdot \mathbf{u} = 0 \} \quad (4.36)$$

consisting of all vectors orthogonal to  $\mathbf{u}$ , as illustrated in Figure 4.3. It is a symmetric orthogonal matrix, and so

$$H^T = H, \quad H^2 = \mathbf{I}, \quad H^{-1} = H. \quad (4.37)$$



**Figure 4.3.** Elementary Reflection Matrix.

The proof is straightforward: symmetry is immediate, while

$$HH^T = H^2 = (I - 2\mathbf{u}\mathbf{u}^T)(I - 2\mathbf{u}\mathbf{u}^T) = I - 4\mathbf{u}\mathbf{u}^T + 4\mathbf{u}(\mathbf{u}^T\mathbf{u})\mathbf{u}^T = I,$$

since, by assumption,  $\mathbf{u}^T\mathbf{u} = \|\mathbf{u}\|^2 = 1$ . Thus, by suitably forming the unit vector  $\mathbf{u}$ , we can construct a Householder matrix that interchanges any two vectors of the same length.

**Lemma 4.28.** Let  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$  with  $\|\mathbf{v}\| = \|\mathbf{w}\|$ . Set  $\mathbf{u} = (\mathbf{v} - \mathbf{w})/\|\mathbf{v} - \mathbf{w}\|$ . Let  $H = I - 2\mathbf{u}\mathbf{u}^T$  be the corresponding elementary reflection matrix. Then  $H\mathbf{v} = \mathbf{w}$  and  $H\mathbf{w} = \mathbf{v}$ .

*Proof:* Keeping in mind that  $\mathbf{v}$  and  $\mathbf{w}$  have the same Euclidean norm, we compute

$$\begin{aligned} H\mathbf{v} &= (I - 2\mathbf{u}\mathbf{u}^T)\mathbf{v} = \mathbf{v} - 2 \frac{(\mathbf{v} - \mathbf{w})(\mathbf{v} - \mathbf{w})^T\mathbf{v}}{\|\mathbf{v} - \mathbf{w}\|^2} \\ &= \mathbf{v} - 2 \frac{\|\mathbf{v}\|^2 - \mathbf{w} \cdot \mathbf{v}}{2\|\mathbf{v}\|^2 - 2\mathbf{v} \cdot \mathbf{w}} (\mathbf{v} - \mathbf{w}) = \mathbf{v} - (\mathbf{v} - \mathbf{w}) = \mathbf{w}. \end{aligned}$$

The proof of the second equation is similar. *Q.E.D.*

In the first phase of Householder's method, we introduce the elementary reflection matrix that maps the first column  $\mathbf{v}_1$  of the matrix  $A$  to a multiple of the first standard basis vector, namely  $\mathbf{w}_1 = \|\mathbf{v}_1\| \mathbf{e}_1$ , noting that  $\|\mathbf{v}_1\| = \|\mathbf{w}_1\|$ . Assuming  $\mathbf{v}_1 \neq c\mathbf{e}_1$ , we define the first unit vector and corresponding elementary reflection matrix as

$$\mathbf{u}_1 = \frac{\mathbf{v}_1 - \|\mathbf{v}_1\| \mathbf{e}_1}{\|\mathbf{v}_1 - \|\mathbf{v}_1\| \mathbf{e}_1\|}, \quad H_1 = I - 2\mathbf{u}_1\mathbf{u}_1^T.$$

On the other hand, if  $\mathbf{v}_1 = c\mathbf{e}_1$  is already in the desired form, then we set  $\mathbf{u}_1 = \mathbf{0}$  and  $H_1 = I$ . Since, by the lemma,  $H_1\mathbf{v}_1 = \mathbf{w}_1$ , when we multiply  $A$  on the left by  $H_1$ , we obtain a matrix

$$A_2 = H_1 A = \begin{pmatrix} r_{11} & \tilde{a}_{12} & \tilde{a}_{13} & \dots & \tilde{a}_{1n} \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \dots & \tilde{a}_{2n} \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \dots & \tilde{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{n2} & \tilde{a}_{n3} & \dots & \tilde{a}_{nn} \end{pmatrix}$$

whose first column is in the desired upper triangular form.

In the next phase, we construct a second elementary reflection matrix to make all the entries below the diagonal in the second column of  $A_2$  zero, keeping in mind that, at the

same time, we should not mess up the first column. The latter requirement tells us that the vector used for the reflection should have a zero in its first entry. The correct choice is to set

$$\tilde{\mathbf{v}}_2 = (0, \tilde{a}_{22}, \tilde{a}_{32}, \dots, \tilde{a}_{n2})^T, \quad \mathbf{u}_2 = \frac{\tilde{\mathbf{v}}_2 - \|\tilde{\mathbf{v}}_2\| \mathbf{e}_2}{\|\tilde{\mathbf{v}}_2 - \|\tilde{\mathbf{v}}_2\| \mathbf{e}_2\|}, \quad H_2 = \mathbf{I} - 2 \mathbf{u}_2 \mathbf{u}_2^T.$$

As before, if  $\tilde{\mathbf{v}}_2 = c \mathbf{e}_2$ , then  $\mathbf{u}_2 = \mathbf{0}$  and  $H_2 = \mathbf{I}$ . The net effect is

$$A_3 = H_2 A_2 = \begin{pmatrix} r_{11} & r_{12} & \hat{a}_{13} & \dots & \hat{a}_{1n} \\ 0 & r_{22} & \hat{a}_{23} & \dots & \hat{a}_{2n} \\ 0 & 0 & \hat{a}_{33} & \dots & \hat{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \hat{a}_{n3} & \dots & \hat{a}_{nn} \end{pmatrix},$$

and now the first two columns are in upper triangular form.

The process continues; at the  $k^{\text{th}}$  stage, we are dealing with a matrix  $A_k$  whose first  $k-1$  columns coincide with the first  $k$  columns of the eventual upper triangular matrix  $R$ . Let  $\hat{\mathbf{v}}_k$  denote the vector obtained from the  $k^{\text{th}}$  column of  $A_k$  by setting its initial  $k-1$  entries equal to 0. We define the  $k^{\text{th}}$  Householder vector and corresponding elementary reflection matrix by

$$\mathbf{w}_k = \hat{\mathbf{v}}_k - \|\hat{\mathbf{v}}_k\| \mathbf{e}_k, \quad \mathbf{u}_k = \begin{cases} \mathbf{w}_k / \|\mathbf{w}_k\|, & \text{if } \mathbf{w}_k \neq \mathbf{0}, \\ \mathbf{0}, & \text{if } \mathbf{w}_k = \mathbf{0}, \end{cases} \quad (4.38)$$

$$H_k = \mathbf{I} - 2 \mathbf{u}_k \mathbf{u}_k^T, \quad A_{k+1} = H_k A_k.$$

The process is completed after  $n-1$  steps, and the final result is

$$R = H_{n-1} A_{n-1} = H_{n-1} H_{n-2} \cdots H_1 A = Q^T A, \quad \text{where} \quad Q = H_1 H_2 \cdots H_{n-1}$$

is an orthogonal matrix, since it is the product of orthogonal matrices, cf. Proposition 4.23. In this manner, we have reproduced a<sup>†</sup>  $Q R$  factorization of

$$A = Q R = H_1 H_2 \cdots H_{n-1} R. \quad (4.39)$$

**Example 4.29.** Let us implement Householder's Method on the particular matrix

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \\ -1 & 2 & 3 \end{pmatrix}$$

considered earlier in Example 4.25. The first Householder vector

$$\hat{\mathbf{v}}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} - \sqrt{3} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -.7321 \\ 1 \\ -1 \end{pmatrix}$$

leads to the elementary reflection matrix

$$H_1 = \begin{pmatrix} .5774 & .5774 & -.5774 \\ .5774 & .2113 & .7887 \\ -.5774 & .7887 & .2113 \end{pmatrix}, \quad \text{whereby } A_2 = H_1 A = \begin{pmatrix} 1.7321 & -.5774 & -1.7321 \\ 0 & 2.1547 & 3.0981 \\ 0 & -.1547 & -2.0981 \end{pmatrix}.$$

---

<sup>†</sup> The upper triangular matrix  $R$  may not have positive diagonal entries; if desired, this can be easily fixed by changing the signs of the appropriate columns of  $Q$ .

To construct the second and final Householder matrix, we start with the second column of  $A_2$  and then set the first entry to 0; the resulting Householder vector is

$$\hat{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 2.1547 \\ -.1547 \end{pmatrix} - 2.1603 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -.0055 \\ -.1547 \end{pmatrix}.$$

Therefore,

$$H_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & .9974 & -.0716 \\ 0 & -.0716 & -.9974 \end{pmatrix}, \quad \text{and so} \quad R = H_2 A_2 = \begin{pmatrix} 1.7321 & -.5774 & -1.7321 \\ 0 & 2.1603 & 3.2404 \\ 0 & 0 & 1.8708 \end{pmatrix}$$

is the upper triangular matrix in the  $QR$  decomposition of  $A$ . The orthogonal matrix  $Q$  is obtained by multiplying the reflection matrices:

$$Q = H_1 H_2 = \begin{pmatrix} .5774 & .6172 & .5345 \\ .5774 & .1543 & -.8018 \\ -.5774 & .7715 & -.2673 \end{pmatrix},$$

which numerically reconfirms the previous factorization (4.33).

**Remark.** If the purpose of the  $QR$  factorization is to solve a linear system via (4.34), it is not necessary to explicitly multiply out the Householder matrices to form  $Q$ ; we merely need to store the corresponding unit Householder vectors  $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$ . The solution to

$$A \mathbf{x} = Q R \mathbf{x} = \mathbf{b} \quad \text{can be found by solving} \quad R \mathbf{x} = H_{n-1} H_{n-2} \cdots H_1 \mathbf{b} \quad (4.40)$$

by Back Substitution. This is the method of choice for moderately ill-conditioned systems. Severe ill-conditioning will defeat even this ingenious approach, and accurately solving such systems can be an extreme challenge.

## Exercises

4.3.26. Write down the  $QR$  matrix factorization corresponding to the vectors in Example 4.17.

4.3.27. Find the  $QR$  factorization of the following matrices: (a)  $\begin{pmatrix} 1 & -3 \\ 2 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 4 & 3 \\ 3 & 2 \end{pmatrix}$ , (c)  $\begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & 3 \\ -1 & -1 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 0 & 1 & 2 \\ -1 & 1 & 1 \\ -1 & 1 & 3 \end{pmatrix}$ , (e)  $\begin{pmatrix} 0 & 0 & 2 \\ 0 & 4 & 1 \\ -1 & 0 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$ .

4.3.28. For each of the following linear systems, find the  $QR$  factorization of the coefficient

matrix, and then use your factorization to solve the system: (i)  $\begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ ,  
(ii)  $\begin{pmatrix} 2 & 1 & -1 \\ 1 & 0 & 2 \\ 2 & -1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ .

♣ 4.3.29. Use the numerically stable version of the Gram–Schmidt process to find the  $QR$  factorizations of the  $3 \times 3$ ,  $4 \times 4$  and  $5 \times 5$  versions of the tridiagonal matrix that has 4's along the diagonal and 1's on the sub- and super-diagonals, as in Example 1.37.

◇ 4.3.30. Prove that the  $QR$  factorization of a matrix is unique if all the diagonal entries of  $R$  are assumed to be positive. *Hint:* Use Exercise 4.3.12.

◇ 4.3.31. (a) How many arithmetic operations are required to compute the  $QR$  factorization of an  $n \times n$  matrix? (b) How many additional operations are needed to utilize the factorization to solve a linear system  $A\mathbf{x} = \mathbf{b}$  via (4.34)? (c) Compare the amount of computational effort with standard Gaussian Elimination.

◇ 4.3.32. Suppose  $A$  is an  $m \times n$  matrix with  $\text{rank } A = n$ . (a) Show that applying the Gram–Schmidt algorithm to the columns of  $A$  produces an orthonormal basis for  $\text{img } A$ . (b) Prove that this is equivalent to the matrix factorization  $A = QR$ , where  $Q$  is an  $m \times n$  matrix with orthonormal columns, while  $R$  is a nonsingular  $n \times n$  upper triangular matrix. (c) Show that the  $QR$  program in the text also works for rectangular,  $m \times n$ , matrices as stated, the only modification being that the row indices  $i$  run from 1 to  $m$ . (d) Apply this method to factor

$$(i) \begin{pmatrix} 1 & -1 \\ 2 & 3 \\ 0 & 2 \end{pmatrix}, \quad (ii) \begin{pmatrix} -3 & 2 \\ 1 & -1 \\ 4 & 1 \end{pmatrix}, \quad (iii) \begin{pmatrix} -1 & 1 \\ 1 & -2 \\ -1 & -3 \\ 0 & 5 \end{pmatrix}, \quad (iv) \begin{pmatrix} 0 & 1 & 2 \\ -3 & 1 & -1 \\ -1 & 0 & -2 \\ 1 & 1 & -2 \end{pmatrix}.$$

(e) Explain what happens if  $\text{rank } A < n$ .

◇ 4.3.33. (a) According to Exercise 4.2.14, the Gram–Schmidt process can also be applied to produce orthonormal bases of complex vector spaces. In the case of  $\mathbb{C}^n$ , explain how this is equivalent to the factorization of a nonsingular complex matrix  $A = UR$  into the product of a unitary matrix  $U$  (see Exercise 4.3.25) and a nonsingular upper triangular matrix  $R$ .

(b) Factor the following complex matrices into unitary times upper triangular:

$$(i) \begin{pmatrix} i & 1 \\ -1 & 2i \end{pmatrix}, \quad (ii) \begin{pmatrix} 1+i & 2-i \\ 1-i & -i \end{pmatrix}, \quad (iii) \begin{pmatrix} i & 1 & 0 \\ 1 & i & 1 \\ 0 & 1 & i \end{pmatrix}, \quad (iv) \begin{pmatrix} i & 1 & -i \\ 1-i & 0 & 1+i \\ -1 & 2+3i & 1 \end{pmatrix}.$$

(c) What can you say about uniqueness of the factorization?

4.3.34. (a) Write down the Householder matrices corresponding to the following unit vectors:

(i)  $(1, 0)^T$ , (ii)  $(\frac{3}{5}, \frac{4}{5})^T$ , (iii)  $(0, 1, 0)^T$ , (iv)  $(\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}})^T$ . (b) Find all vectors fixed by a Householder matrix, i.e.,  $H\mathbf{v} = \mathbf{v}$  — first for the matrices in part (a), and then in general. (c) Is a Householder matrix a proper or improper orthogonal matrix?

4.3.35. Use Householder's Method to solve Exercises 4.3.27 and 4.3.29.

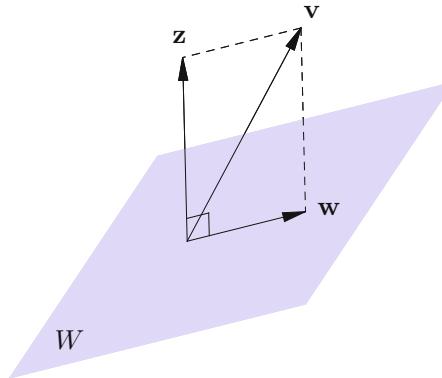
♣ 4.3.36. Let  $H_n = Q_n R_n$  be the  $QR$  factorization of the  $n \times n$  Hilbert matrix (1.72). (a) Find  $Q_n$  and  $R_n$  for  $n = 2, 3, 4$ . (b) Use a computer to find  $Q_n$  and  $R_n$  for  $n = 10$  and  $20$ .

(c) Let  $\mathbf{x}^* \in \mathbb{R}^n$  denote the vector whose  $i^{\text{th}}$  entry is  $x_i^* = (-1)^i i/(i+1)$ . For the values of  $n$  in parts (a) and (b), compute  $\mathbf{y}^* = H_n \mathbf{x}^*$ . Then solve the system  $H_n \mathbf{x} = \mathbf{y}^*$  (i) directly using Gaussian Elimination; (ii) using the  $QR$  factorization based on (4.34); (iii) using Householder's Method. Compare the results to the correct solution  $\mathbf{x}^*$  and discuss the pros and cons of each method.

4.3.37. Write out a pseudocode program to implement Householder's Method. The input should be an  $n \times n$  matrix  $A$  and the output should be the Householder unit vectors  $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$  and the upper triangular matrix  $R$ . Test your code on one of the examples in Exercises 4.3.26–28.

## 4.4 Orthogonal Projections and Orthogonal Subspaces

Orthogonality is important, not just for individual vectors, but also for subspaces. In this section, we develop two concepts. First, we investigate the orthogonal projection of a vector onto a subspace, an operation that plays a key role in least squares minimization and data



**Figure 4.4.** The Orthogonal Projection of a Vector onto a Subspace.

fitting, as we shall discuss in Chapter 5. Second, we develop the concept of orthogonality for a pair of subspaces, culminating with a proof of the orthogonality of the fundamental subspaces associated with an  $m \times n$  matrix that at last reveals the striking geometry that underlies linear systems of equations and matrix multiplication.

### Orthogonal Projection

Throughout this section,  $W \subset V$  will be a finite-dimensional subspace of a real inner product space. The inner product space  $V$  is allowed to be infinite-dimensional. But, to facilitate your geometric intuition, you may initially want to view  $W$  as a subspace of Euclidean space  $V = \mathbb{R}^m$  equipped with the ordinary dot product.

**Definition 4.30.** A vector  $\mathbf{z} \in V$  is said to be *orthogonal* to the subspace  $W \subset V$  if it is orthogonal to every vector in  $W$ , so  $\langle \mathbf{z}, \mathbf{w} \rangle = 0$  for all  $\mathbf{w} \in W$ .

Given a basis  $\mathbf{w}_1, \dots, \mathbf{w}_n$  of the subspace  $W$ , we note that  $\mathbf{z}$  is orthogonal to  $W$  if and only if it is orthogonal to every basis vector:  $\langle \mathbf{z}, \mathbf{w}_i \rangle = 0$  for  $i = 1, \dots, n$ . Indeed, any other vector in  $W$  has the form  $\mathbf{w} = c_1 \mathbf{w}_1 + \dots + c_n \mathbf{w}_n$ , and hence, by linearity,  $\langle \mathbf{z}, \mathbf{w} \rangle = c_1 \langle \mathbf{z}, \mathbf{w}_1 \rangle + \dots + c_n \langle \mathbf{z}, \mathbf{w}_n \rangle = 0$ , as required.

**Definition 4.31.** The *orthogonal projection* of  $\mathbf{v}$  onto the subspace  $W$  is the element  $\mathbf{w} \in W$  that makes the difference  $\mathbf{z} = \mathbf{v} - \mathbf{w}$  orthogonal to  $W$ .

The geometric configuration underlying orthogonal projection is sketched in [Figure 4.4](#). As we shall see, the orthogonal projection is unique. Note that  $\mathbf{v} = \mathbf{w} + \mathbf{z}$  is the sum of its orthogonal projection  $\mathbf{w} \in V$  and the perpendicular vector  $\mathbf{z} \perp W$ .

The explicit construction is greatly simplified by taking an orthonormal basis of the subspace, which, if necessary, can be arranged by applying the Gram–Schmidt process to a known basis. (The direct construction of the orthogonal projection in terms of a non-orthogonal basis appears in [Exercise 4.4.10](#).)

**Theorem 4.32.** Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be an orthonormal basis for the subspace  $W \subset V$ . Then the orthogonal projection of  $\mathbf{v} \in V$  onto  $W$  is given by

$$\mathbf{w} = c_1 \mathbf{u}_1 + \dots + c_n \mathbf{u}_n \quad \text{where} \quad c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle, \quad i = 1, \dots, n. \quad (4.41)$$

*Proof:* First, since  $\mathbf{u}_1, \dots, \mathbf{u}_n$  form a basis of the subspace, the orthogonal projection element must be some linear combination thereof:  $\mathbf{w} = c_1 \mathbf{u}_1 + \dots + c_n \mathbf{u}_n$ . Definition 4.31 requires that the difference  $\mathbf{z} = \mathbf{v} - \mathbf{w}$  be orthogonal to  $W$ , and, as noted above, it suffices to check orthogonality to the basis vectors. By our orthonormality assumption,

$$\begin{aligned} 0 &= \langle \mathbf{z}, \mathbf{u}_i \rangle = \langle \mathbf{v} - \mathbf{w}, \mathbf{u}_i \rangle = \langle \mathbf{v} - c_1 \mathbf{u}_1 - \dots - c_n \mathbf{u}_n, \mathbf{u}_i \rangle \\ &= \langle \mathbf{v}, \mathbf{u}_i \rangle - c_1 \langle \mathbf{u}_1, \mathbf{u}_i \rangle - \dots - c_n \langle \mathbf{u}_n, \mathbf{u}_i \rangle = \langle \mathbf{v}, \mathbf{u}_i \rangle - c_i. \end{aligned}$$

The coefficients  $c_i = \langle \mathbf{v}, \mathbf{u}_i \rangle$  of the orthogonal projection  $\mathbf{w}$  are thus uniquely prescribed by the orthogonality requirement, which thereby proves its uniqueness. *Q.E.D.*

More generally, if we employ an orthogonal basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  for the subspace  $W$ , then the same argument demonstrates that the orthogonal projection of  $\mathbf{v}$  onto  $W$  is given by

$$\mathbf{w} = a_1 \mathbf{v}_1 + \dots + a_n \mathbf{v}_n, \quad \text{where} \quad a_i = \frac{\langle \mathbf{v}, \mathbf{v}_i \rangle}{\|\mathbf{v}_i\|^2}, \quad i = 1, \dots, n. \quad (4.42)$$

We could equally well replace the orthogonal basis by the orthonormal basis obtained by dividing each vector by its length:  $\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$ . The reader should be able to prove that the two formulas (4.41, 42) for the orthogonal projection yield the same vector  $\mathbf{w}$ .

**Example 4.33.** Consider the plane  $W \subset \mathbb{R}^3$  spanned by the orthogonal vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

According to formula (4.42), the orthogonal projection of  $\mathbf{v} = (1, 0, 0)^T$  onto  $W$  is

$$\mathbf{w} = \frac{\langle \mathbf{v}, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 + \frac{\langle \mathbf{v}, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 = \frac{1}{6} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}.$$

Alternatively, we can replace  $\mathbf{v}_1, \mathbf{v}_2$  by the orthonormal basis

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}, \quad \mathbf{u}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}.$$

Then, using the orthonormal version (4.41),

$$\mathbf{w} = \langle \mathbf{v}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{v}, \mathbf{u}_2 \rangle \mathbf{u}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} + \frac{1}{\sqrt{3}} \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}.$$

The answer is, of course, the same. As the reader may notice, while the theoretical formula is simpler when written in an orthonormal basis, for hand computations the orthogonal basis version avoids having to deal with square roots. (Of course, when the numerical computation is performed on a computer, this is not a significant issue.)

An intriguing observation is that the coefficients in the orthogonal projection formulas (4.41–42) coincide with the formulas (4.4, 7) for writing a vector in terms of an orthonormal

or orthogonal basis. Indeed, if  $\mathbf{v}$  were an element of  $W$ , then it would coincide with its orthogonal projection,  $\mathbf{w} = \mathbf{v}$ . (Why?) As a result, the orthogonal projection formula include the orthogonal basis formula as a special case.

It is also worth noting that the *same* formulae occur in the Gram–Schmidt algorithm, cf. (4.19). This observation leads to a useful geometric interpretation of the Gram–Schmidt construction. For each  $k = 1, \dots, n$ , let

$$W_k = \text{span} \{ \mathbf{w}_1, \dots, \mathbf{w}_k \} = \text{span} \{ \mathbf{v}_1, \dots, \mathbf{v}_k \} = \text{span} \{ \mathbf{u}_1, \dots, \mathbf{u}_k \} \quad (4.43)$$

denote the  $k$ -dimensional subspace spanned by the first  $k$  basis elements, which is the same as that spanned by their orthogonalized and orthonormalized counterparts. In view of (4.41), the basic Gram–Schmidt formula (4.19) can be re-expressed in the form  $\mathbf{v}_k = \mathbf{w}_k - \mathbf{p}_k$ , where  $\mathbf{p}_k$  is the orthogonal projection of  $\mathbf{w}_k$  onto the subspace  $W_{k-1}$ . The resulting vector  $\mathbf{v}_k$  is, by construction, orthogonal to the subspace, and hence orthogonal to all of the previous basis elements, which serves to rejustify the Gram–Schmidt construction.

## Exercises

*Note:* Use the dot product and Euclidean norm unless otherwise specified.

4.4.1. Determine which of the vectors  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} -2 \\ 2 \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 2 \\ -1 \\ -3 \end{pmatrix}$ ,  $\mathbf{v}_4 = \begin{pmatrix} -1 \\ 3 \\ 4 \end{pmatrix}$ , is

- orthogonal to (a) the line spanned by  $\begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$ ; (b) the plane spanned by  $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ ;  
 (c) the plane defined by  $x - y - z = 0$ ; (d) the kernel of the matrix  $\begin{pmatrix} 1 & -1 & -1 \\ 3 & -2 & -4 \end{pmatrix}$ ;  
 (e) the image of the matrix  $\begin{pmatrix} -3 & 1 \\ 3 & -1 \\ -1 & 0 \end{pmatrix}$ ; (f) the cokernel of the matrix  $\begin{pmatrix} -1 & 0 & 3 \\ 2 & 1 & -2 \\ 3 & 1 & -5 \end{pmatrix}$ .

4.4.2. Find the orthogonal projection of the vector  $\mathbf{v} = (1, 1, 1)^T$  onto the following subspaces, using the indicated orthonormal/orthogonal bases: (a) the line in the direction

- $\left( -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)^T$ ; (b) the line spanned by  $(2, -1, 3)^T$ ; (c) the plane spanned by  $(1, 1, 0)^T$ ,  $(-2, 2, 1)^T$ ; (d) the plane spanned by  $\left( -\frac{3}{5}, \frac{4}{5}, 0 \right)^T$ ,  $\left( \frac{4}{13}, \frac{3}{13}, -\frac{12}{13} \right)^T$ .

4.4.3. Find the orthogonal projection of  $\mathbf{v} = (1, 2, -1, 2)^T$  onto the following subspaces:

- (a) the span of  $\begin{pmatrix} 1 \\ -1 \\ 2 \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 2 \\ 1 \\ 0 \\ -1 \end{pmatrix}$ ; (b) the image of the matrix  $\begin{pmatrix} 1 & 2 \\ -1 & 1 \\ 0 & 3 \\ -1 & 1 \end{pmatrix}$ ; (c) the kernel of the matrix  $\begin{pmatrix} 1 & -1 & 0 & 1 \\ -2 & 1 & 1 & 0 \end{pmatrix}$ ; (d) the subspace orthogonal to  $\mathbf{a} = (1, -1, 0, 1)^T$ .

**Warning.** Make sure you have an orthogonal basis before applying formula (4.42)!

4.4.4. Find the orthogonal projection of the vector  $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$  onto the image of  $\begin{pmatrix} 3 & 2 \\ 2 & -2 \\ 1 & -2 \end{pmatrix}$ .

4.4.5. Find the orthogonal projection of the vector  $\mathbf{v} = (1, 3, -1)^T$  onto the plane spanned by  $(-1, 2, 1)^T$ ,  $(2, 1, -3)^T$  by first using the Gram–Schmidt process to construct an orthogonal basis.

4.4.6. Find the orthogonal projection of  $\mathbf{v} = (1, 2, -1, 2)^T$  onto the span of  $(1, -1, 2, 5)^T$  and  $(2, 1, 0, -1)^T$  using the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 4v_1 w_1 + 3v_2 w_2 + 2v_3 w_3 + v_4 w_4$ .

4.4.7. Redo Exercise 4.4.2 using

(i) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1 w_1 + 2v_2 w_2 + v_3 w_3$ ;

(ii) the inner product induced by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ .

4.4.8. (a) Prove that the set of all vectors orthogonal to a given subspace  $V \subset \mathbb{R}^m$  forms a subspace. (b) Find a basis for the set of all vectors in  $\mathbb{R}^4$  that are orthogonal to the subspace spanned by  $(1, 2, 0, -1)^T, (2, 0, 3, 1)^T$ .

◇ 4.4.9. Let  $\mathbf{u}_1, \dots, \mathbf{u}_k$  be an orthonormal basis for the subspace  $W \subset \mathbb{R}^m$ . Let

$A = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_k)$  be the  $m \times k$  matrix whose columns are the orthonormal basis vectors, and define  $P = AA^T$  to be the corresponding *projection matrix*. (a) Given  $\mathbf{v} \in \mathbb{R}^n$ , prove that its orthogonal projection  $\mathbf{w} \in W$  is given by matrix multiplication:  $\mathbf{w} = P\mathbf{v}$ .

(b) Prove that  $P = P^T$  is symmetric. (c) Prove that  $P$  is idempotent:  $P^2 = P$ . Give a geometrical explanation of this fact. (d) Prove that  $\text{rank } P = k$ . (e) Write out the projection matrix corresponding to the subspaces spanned by

$$(i) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad (ii) \begin{pmatrix} \frac{2}{3} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}, \quad (iii) \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}, \quad (iv) \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}.$$

◇ 4.4.10. Let  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be an arbitrary basis of the subspace  $W \subset \mathbb{R}^m$ . Let  $A = (\mathbf{w}_1, \dots, \mathbf{w}_n)$  be the  $m \times n$  matrix whose columns are the basis vectors, so that  $W = \text{img } A$  and  $\text{rank } A = n$ . (a) Prove that the corresponding *projection matrix*  $P = A(A^T A)^{-1} A^T$  is idempotent:  $P^2 = P$ . (b) Prove that  $P$  is symmetric. (c) Prove that  $\text{img } P = W$ . (d) (e) Prove that the orthogonal projection of  $\mathbf{v} \in \mathbb{R}^n$  onto  $\mathbf{w} \in W$  is obtained by multiplying by the projection matrix:  $\mathbf{w} = P\mathbf{v}$ . (f) Show that if  $A$  is nonsingular, then  $P = I$ . How do you interpret this in light of part (e)? (g) Explain why Exercise 4.4.9 is a special case of this result. (h) Show that if  $A = QR$  is the factorization of  $A$  given in Exercise 4.3.32, then  $P = QQ^T$ . Why is  $P \neq I$ ?

4.4.11. Use the projection matrix method of Exercise 4.4.10 to find the orthogonal projection of  $\mathbf{v} = (1, 0, 0, 0)^T$  onto the image of the following matrices:

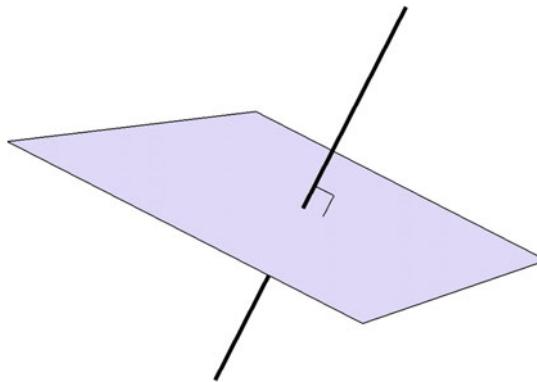
$$(a) \begin{pmatrix} 5 \\ -5 \\ -7 \\ 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 0 \\ -1 & 2 \\ 0 & -1 \\ 1 & 2 \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & -1 \\ -3 & 1 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad (d) \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 2 \\ 1 & 1 & 1 \\ -2 & -1 & 0 \end{pmatrix}.$$

## Orthogonal Subspaces

We now extend the notion of orthogonality from individual elements to entire subspaces of an inner product space  $V$ .

**Definition 4.34.** Two subspaces  $W, Z \subset V$  are called *orthogonal* if every vector in  $W$  is orthogonal to every vector in  $Z$ .

In other words,  $W$  and  $Z$  are orthogonal subspaces if and only if  $\langle \mathbf{w}, \mathbf{z} \rangle = 0$  for every  $\mathbf{w} \in W$  and  $\mathbf{z} \in Z$ . In practice, one only needs to check orthogonality of basis elements,



**Figure 4.5.** Orthogonal Complement to a Line.

or, more generally, spanning sets.

**Lemma 4.35.** If  $\mathbf{w}_1, \dots, \mathbf{w}_k$  span  $W$  and  $\mathbf{z}_1, \dots, \mathbf{z}_l$  span  $Z$ , then  $W$  and  $Z$  are orthogonal subspaces if and only if  $\langle \mathbf{w}_i, \mathbf{z}_j \rangle = 0$  for all  $i = 1, \dots, k$  and  $j = 1, \dots, l$ .

The proof of this lemma is left to the reader; see Exercise 4.4.26.

**Example 4.36.** Let  $V = \mathbb{R}^3$  have the ordinary dot product. Then the plane  $W \subset \mathbb{R}^3$  defined by the equation  $2x - y + 3z = 0$  is orthogonal to the line  $Z$  spanned by its normal vector  $\mathbf{n} = (2, -1, 3)^T$ . Indeed, every  $\mathbf{w} = (x, y, z)^T \in W$  satisfies the orthogonality condition  $\mathbf{w} \cdot \mathbf{n} = 2x - y + 3z = 0$ , which is simply the equation for the plane.

**Example 4.37.** Let  $W$  be the span of  $\mathbf{w}_1 = (1, -2, 0, 1)^T$ ,  $\mathbf{w}_2 = (3, -5, 2, 1)^T$ , and let  $Z$  be the span of the vectors  $\mathbf{z}_1 = (3, 2, 0, 1)^T$ ,  $\mathbf{z}_2 = (1, 0, -1, -1)^T$ . We find that  $\mathbf{w}_1 \cdot \mathbf{z}_1 = \mathbf{w}_1 \cdot \mathbf{z}_2 = \mathbf{w}_2 \cdot \mathbf{z}_1 = \mathbf{w}_2 \cdot \mathbf{z}_2 = 0$ , and so  $W$  and  $Z$  are orthogonal two-dimensional subspaces of  $\mathbb{R}^4$  under the Euclidean dot product.

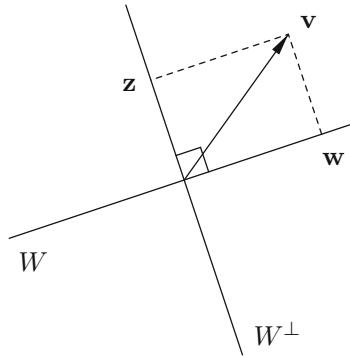
**Definition 4.38.** The *orthogonal complement* of a subspace  $W \subset V$ , denoted<sup>†</sup>  $W^\perp$ , is defined as the set of all vectors that are orthogonal to  $W$ :

$$W^\perp = \{ \mathbf{v} \in V \mid \langle \mathbf{v}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in W \}. \quad (4.44)$$

If  $W$  is the one-dimensional subspace (line) spanned by a single vector  $\mathbf{w} \neq \mathbf{0}$  then we also denote  $W^\perp$  by  $\mathbf{w}^\perp$ , as in (4.36). One easily checks that the orthogonal complement  $W^\perp$  is also a subspace. Moreover,  $W \cap W^\perp = \{\mathbf{0}\}$ . (Why?) Keep in mind that the orthogonal complement will depend upon which inner product is being used.

**Example 4.39.** Let  $W = \{ (t, 2t, 3t)^T \mid t \in \mathbb{R} \}$  be the line (one-dimensional subspace) in the direction of the vector  $\mathbf{w}_1 = (1, 2, 3)^T \in \mathbb{R}^3$ . Under the dot product, its orthogonal

<sup>†</sup> And usually pronounced “W perp”



**Figure 4.6.** Orthogonal Decomposition of a Vector.

complement  $W^\perp = \mathbf{w}_1^\perp$  is the plane passing through the origin having normal vector  $\mathbf{w}_1$ , as sketched in Figure 4.5. In other words,  $\mathbf{z} = (x, y, z)^T \in W^\perp$  if and only if

$$\mathbf{z} \cdot \mathbf{w}_1 = x + 2y + 3z = 0. \quad (4.45)$$

Thus,  $W^\perp$  is characterized as the solution space of the homogeneous linear equation (4.45), or, equivalently, the kernel of the  $1 \times 3$  matrix  $A = \mathbf{w}_1^T = (1 \ 2 \ 3)$ . We can write the general solution in the form

$$\mathbf{z} = \begin{pmatrix} -2y - 3z \\ y \\ z \end{pmatrix} = y \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} = y \mathbf{z}_1 + z \mathbf{z}_2,$$

where  $y, z$  are the free variables. The indicated vectors  $\mathbf{z}_1 = (-2, 1, 0)^T$ ,  $\mathbf{z}_2 = (-3, 0, 1)^T$ , form a (non-orthogonal) basis for the orthogonal complement  $W^\perp$ .

**Proposition 4.40.** Suppose that  $W \subset V$  is a finite-dimensional subspace of an inner product space. Then every vector  $\mathbf{v} \in V$  can be uniquely decomposed into  $\mathbf{v} = \mathbf{w} + \mathbf{z}$ , where  $\mathbf{w} \in W$  and  $\mathbf{z} \in W^\perp$ .

*Proof:* We let  $\mathbf{w} \in W$  be the orthogonal projection of  $\mathbf{v}$  onto  $W$ . Then  $\mathbf{z} = \mathbf{v} - \mathbf{w}$  is, by definition, orthogonal to  $W$  and hence belongs to  $W^\perp$ . Note that  $\mathbf{z}$  can be viewed as the orthogonal projection of  $\mathbf{v}$  onto the complementary subspace  $W^\perp$  (provided it is finite-dimensional). If we are given two such decompositions,  $\mathbf{v} = \mathbf{w} + \mathbf{z} = \tilde{\mathbf{w}} + \tilde{\mathbf{z}}$ , then  $\mathbf{w} - \tilde{\mathbf{w}} = \tilde{\mathbf{z}} - \mathbf{z}$ . The left-hand side of this equation lies in  $W$ , while the right-hand side belongs to  $W^\perp$ . But, as we already noted, the only vector that belongs to both  $W$  and  $W^\perp$  is the zero vector. Thus,  $\mathbf{w} - \tilde{\mathbf{w}} = \mathbf{0} = \tilde{\mathbf{z}} - \mathbf{z}$ , so  $\mathbf{w} = \tilde{\mathbf{w}}$  and  $\mathbf{z} = \tilde{\mathbf{z}}$ , which proves uniqueness. *Q.E.D.*

As a direct consequence of Exercise 2.4.26, in a finite-dimensional inner product space, a subspace and its orthogonal complement have complementary dimensions:

**Proposition 4.41.** If  $W \subset V$  is a subspace with  $\dim W = n$  and  $\dim V = m$ , then  $\dim W^\perp = m - n$ .

**Example 4.42.** Return to the situation described in Example 4.39. Let us decompose the vector  $\mathbf{v} = (1, 0, 0)^T \in \mathbb{R}^3$  into a sum  $\mathbf{v} = \mathbf{w} + \mathbf{z}$  of a vector  $\mathbf{w}$  lying on the line  $W$

and a vector  $\mathbf{z}$  belonging to its orthogonal plane  $W^\perp$ , defined by (4.45). Each is obtained by an orthogonal projection onto the subspace in question, but we only need to compute one of the two directly, since the second can be obtained by subtracting the first from  $\mathbf{v}$ .

Orthogonal projection onto a one-dimensional subspace is easy, since every basis is, trivially, an orthogonal basis. Thus, the projection of  $\mathbf{v}$  onto the line spanned by

$$\mathbf{w}_1 = (1, 2, 3)^T \quad \text{is} \quad \mathbf{w} = \frac{\langle \mathbf{v}, \mathbf{w}_1 \rangle}{\| \mathbf{w}_1 \|^2} \mathbf{w}_1 = \left( \frac{1}{14}, \frac{2}{14}, \frac{3}{14} \right)^T.$$

The component in  $W^\perp$  is then obtained by subtraction:

$$\mathbf{z} = \mathbf{v} - \mathbf{w} = \left( \frac{13}{14}, -\frac{2}{14}, -\frac{3}{14} \right)^T.$$

Alternatively, one can obtain  $\mathbf{z}$  directly by orthogonal projection onto the plane  $W^\perp$ . But you need to be careful: the basis found in Example 4.39 is not orthogonal, and so you will need to either first convert to an orthogonal basis and then use the orthogonal projection formula (4.42), or apply the more direct result in Exercise 4.4.10.

**Example 4.43.** Let  $W \subset \mathbb{R}^4$  be the two-dimensional subspace spanned by the orthogonal vectors  $\mathbf{w}_1 = (1, 1, 0, 1)^T$  and  $\mathbf{w}_2 = (1, 1, 1, -2)^T$ . Its orthogonal complement  $W^\perp$  (with respect to the Euclidean dot product) is the set of all vectors  $\mathbf{v} = (x, y, z, w)^T$  that satisfy the linear system

$$\mathbf{v} \cdot \mathbf{w}_1 = x + y + w = 0, \quad \mathbf{v} \cdot \mathbf{w}_2 = x + y + z - 2w = 0.$$

Applying the usual algorithm — the free variables are  $y$  and  $w$  — we find that the solution space is spanned by

$$\mathbf{z}_1 = (-1, 1, 0, 0)^T, \quad \mathbf{z}_2 = (-1, 0, 3, 1)^T,$$

which form a non-orthogonal basis for  $W^\perp$ . An orthogonal basis

$$\mathbf{y}_1 = \mathbf{z}_1 = (-1, 1, 0, 0)^T, \quad \mathbf{y}_2 = \mathbf{z}_2 - \frac{1}{2}\mathbf{z}_1 = \left( -\frac{1}{2}, -\frac{1}{2}, 3, 1 \right)^T,$$

for  $W^\perp$  is obtained by a single Gram–Schmidt step. To decompose the vector  $\mathbf{v} = (1, 0, 0, 0)^T = \mathbf{w} + \mathbf{z}$ , say, we compute the two orthogonal projections:

$$\begin{aligned} \mathbf{w} &= \frac{1}{3}\mathbf{w}_1 + \frac{1}{7}\mathbf{w}_2 = \left( \frac{10}{21}, \frac{10}{21}, \frac{1}{7}, \frac{1}{21} \right)^T \in W, \\ \mathbf{z} &= \mathbf{v} - \mathbf{w} = -\frac{1}{2}\mathbf{y}_1 - \frac{1}{21}\mathbf{y}_2 = \left( \frac{11}{21}, -\frac{10}{21}, -\frac{1}{7}, -\frac{1}{21} \right)^T \in W^\perp. \end{aligned}$$

**Proposition 4.44.** If  $W$  is a finite-dimensional subspace of an inner product space, then  $(W^\perp)^\perp = W$ .

This result is a corollary of the orthogonal decomposition derived in Proposition 4.40.

**Warning.** Propositions 4.40 and 4.44 are *not* necessarily true for infinite-dimensional subspaces. If  $\dim W = \infty$ , one can assert only that  $W \subseteq (W^\perp)^\perp$ . For example, it can be shown, [19; Exercise 10.2.D], that on every bounded interval  $[a, b]$  the orthogonal complement of the subspace of all polynomials  $\mathcal{P}^{(\infty)} \subset C^0[a, b]$  with respect to the  $L^2$  inner product is trivial:  $(\mathcal{P}^{(\infty)})^\perp = \{0\}$ . This means that the only continuous function that satisfies

$$\langle x^n, f(x) \rangle = \int_a^b x^n f(x) dx = 0 \quad \text{for all } n = 0, 1, 2, \dots$$

is the zero function  $f(x) \equiv 0$ . But the orthogonal complement of  $\{0\}$  is the entire space, and so  $((\mathcal{P}^{(\infty)})^\perp)^\perp = \mathcal{C}^0[a, b] \neq \mathcal{P}^{(\infty)}$ .

The difference is that, in infinite-dimensional function space, a proper subspace  $W \subsetneq V$  can be *dense*<sup>†</sup>, whereas in finite dimensions, every proper subspace is a “thin” subset that occupies only an infinitesimal fraction of the entire vector space. However, this seeming paradox is, interestingly, the reason behind the success of numerical approximation schemes in function space, such as the finite element method, [81].

## Exercises

*Note:* In Exercises 4.4.12–15, use the dot product.

4.4.12. Find the orthogonal complement  $W^\perp$  of the subspaces  $W \subset \mathbb{R}^3$  spanned by the indicated vectors. What is the dimension of  $W^\perp$  in each case?

$$(a) \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}, (b) \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, (c) \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}, (d) \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}, (e) \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, (f) \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, (g) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, (h) \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

4.4.13. Find a basis for the orthogonal complement of the following subspaces of  $\mathbb{R}^3$ : (a) the plane  $3x + 4y - 5z = 0$ ; (b) the line in the direction  $(-2, 1, 3)^T$ ; (c) the image of the matrix  $\begin{pmatrix} 1 & 2 & -1 & 3 \\ -2 & 0 & 2 & 1 \\ -1 & 2 & 1 & 4 \end{pmatrix}$ ; (d) the cokernel of the same matrix.

4.4.14. Find a basis for the orthogonal complement of the following subspaces of  $\mathbb{R}^4$ : (a) the set of solutions to  $-x + 3y - 2z + w = 0$ ; (b) the subspace spanned by  $(1, 2, -1, 3)^T$ ,  $(-2, 0, 1, -2)^T$ ,  $(-1, 2, 0, 1)^T$ ; (c) the kernel of the matrix in Exercise 4.4.13c; (d) the coimage of the same matrix.

4.4.15. Decompose each of the following vectors with respect to the indicated subspace as

$$\mathbf{v} = \mathbf{w} + \mathbf{z}, \text{ where } \mathbf{w} \in W, \mathbf{z} \in W^\perp. \quad (a) \mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, W = \text{span} \left\{ \begin{pmatrix} -3 \\ 1 \end{pmatrix} \right\};$$

$$(b) \mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, W = \text{span} \left\{ \begin{pmatrix} -3 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 5 \end{pmatrix} \right\}; \quad (c) \mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, W = \ker \begin{pmatrix} 1 & 2 & -1 \\ 2 & 0 & 2 \end{pmatrix};$$

$$(d) \mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, W = \text{img} \begin{pmatrix} 1 & 0 & 1 \\ -2 & -1 & 0 \\ 1 & 3 & -5 \end{pmatrix}; \quad (e) \mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, W = \ker \begin{pmatrix} 1 & 0 & 0 & 2 \\ -2 & -1 & 1 & -3 \end{pmatrix}.$$

4.4.16. Redo Exercise 4.4.12 using the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$  instead of the dot product.

4.4.17. Redo Example 4.4.3 with the dot product replaced by the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3 + 4v_4 w_4$ .

◇ 4.4.18. Prove that the orthogonal complement  $W^\perp$  of a subspace  $W \subset V$  is itself a subspace.

<sup>†</sup> In general, a subset  $W \subset V$  of a normed vector space is *dense* if, for every  $\mathbf{v} \in V$ , and every  $\varepsilon > 0$ , one can find  $\mathbf{w} \in W$  with  $\|\mathbf{v} - \mathbf{w}\| < \varepsilon$ . The Weierstrass Approximation Theorem, [19; Theorem 10.2.2], tells us that the polynomials form a dense subspace of the space of continuous functions, and underlies the proof of the result mentioned in the preceding paragraph.

4.4.19. Let  $V = \mathcal{P}^{(4)}$  denote the space of quartic polynomials, with the  $L^2$  inner product  $\langle p, q \rangle = \int_{-1}^1 p(x) q(x) dx$ . Let  $W = \mathcal{P}^{(2)}$  be the subspace of quadratic polynomials.

- (a) Write down the conditions that a polynomial  $p \in \mathcal{P}^{(4)}$  must satisfy in order to belong to the orthogonal complement  $W^\perp$ . (b) Find a basis for and the dimension of  $W^\perp$ .  
 (c) Find an orthogonal basis for  $W^\perp$ .

4.4.20. Let  $W \subset V$ . Prove that (a)  $W \cap W^\perp = \{0\}$ , (b)  $W \subseteq (W^\perp)^\perp$ .

4.4.21. Let  $V$  be an inner product space. Prove that (a)  $V^\perp = \{\mathbf{0}\}$ , (b)  $\{\mathbf{0}\}^\perp = V$ .

4.4.22. Prove that if  $W_1 \subset W_2$  are finite-dimensional subspaces of an inner product space, then  $W_1^\perp \supset W_2^\perp$ .

4.4.23. (a) Show that if  $W, Z \subset \mathbb{R}^n$  are complementary subspaces, then  $W^\perp$  and  $Z^\perp$  are also complementary subspaces. (b) Sketch a picture illustrating this result when  $W$  and  $Z$  are lines in  $\mathbb{R}^2$ .

4.4.24. Prove that if  $W, Z$  are subspaces of an inner product space, then  $(W+Z)^\perp = W^\perp \cap Z^\perp$ . (See Exercise 2.2.22(b) for the definition of the sum of two subspaces.)

◊ 4.4.25. Fill in the details of the proof of Proposition 4.44.

◊ 4.4.26. Prove Lemma 4.35.

◊ 4.4.27. Let  $W \subset V$  with  $\dim V = n$ . Suppose  $\mathbf{w}_1, \dots, \mathbf{w}_m$  is an orthogonal basis for  $W$  and  $\mathbf{w}_{m+1}, \dots, \mathbf{w}_n$  is an orthogonal basis for  $W^\perp$ . (a) Prove that the combination  $\mathbf{w}_1, \dots, \mathbf{w}_n$  forms an orthogonal basis of  $V$ . (b) Show that if  $\mathbf{v} = c_1 \mathbf{w}_1 + \dots + c_n \mathbf{w}_n$  is any vector in  $V$ , then its orthogonal decomposition  $\mathbf{v} = \mathbf{w} + \mathbf{z}$  is given by  $\mathbf{w} = c_1 \mathbf{w}_1 + \dots + c_m \mathbf{w}_m \in W$  and  $\mathbf{z} = c_{m+1} \mathbf{w}_{m+1} + \dots + c_n \mathbf{w}_n \in W^\perp$ .

◊ 4.4.28. Consider the subspace  $W = \{u(a) = 0 = u(b)\}$  of the vector space  $C^0[a, b]$  with the usual  $L^2$  inner product. (a) Show that  $W$  has a complementary subspace of dimension 2. (b) Prove that there does not exist an orthogonal complement of  $W$ . Thus, an infinite-dimensional subspace may not admit an orthogonal complement!

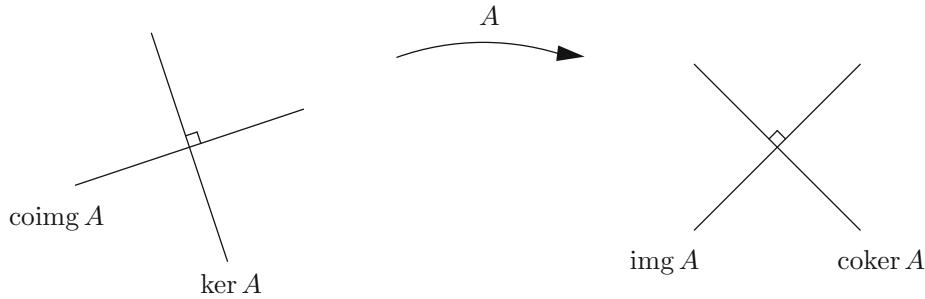
## Orthogonality of the Fundamental Matrix Subspaces and the Fredholm Alternative

In Chapter 2, we introduced the four fundamental subspaces associated with an  $m \times n$  matrix  $A$ . According to the Fundamental Theorem 2.49, the first two, the kernel (null space) and the coimage (row space), are subspaces of  $\mathbb{R}^n$  having complementary dimensions. The second two, the cokernel (left null space) and the image (column space), are subspaces of  $\mathbb{R}^m$ , also of complementary dimensions. In fact, more than this is true — the paired subspaces are orthogonal complements with respect to the standard Euclidean dot product!

**Theorem 4.45.** Let  $A$  be a real  $m \times n$  matrix. Then its kernel and coimage are orthogonal complements as subspaces of  $\mathbb{R}^n$  under the dot product, while its cokernel and image are orthogonal complements in  $\mathbb{R}^m$ , also under the dot product:

$$\ker A = (\text{coimg } A)^\perp \subset \mathbb{R}^n, \quad \text{coker } A = (\text{img } A)^\perp \subset \mathbb{R}^m. \quad (4.46)$$

*Proof:* A vector  $\mathbf{x} \in \mathbb{R}^n$  lies in  $\ker A$  if and only if  $A\mathbf{x} = \mathbf{0}$ . According to the rules of matrix multiplication, the  $i^{\text{th}}$  entry of  $A\mathbf{x}$  equals the vector product of the  $i^{\text{th}}$  row  $\mathbf{r}_i^T$  of



**Figure 4.7.** The Fundamental Matrix Subspaces.

$A$  and  $\mathbf{x}$ . But this product vanishes,  $\mathbf{r}_i^T \mathbf{x} = \mathbf{r}_i \cdot \mathbf{x} = 0$ , if and only if  $\mathbf{x}$  is orthogonal to  $\mathbf{r}_i$ . Therefore,  $\mathbf{x} \in \ker A$  if and only if  $\mathbf{x}$  is orthogonal to all the rows of  $A$ . Since the rows span  $\text{coimg } A$ , this is equivalent to  $\mathbf{x}$  lying in its orthogonal complement  $(\text{coimg } A)^\perp$ , which proves the first statement. Orthogonality of the image and cokernel follows by the same argument applied to the transposed matrix  $A^T$ . *Q.E.D.*

Combining Theorems 2.49 and 4.45, we deduce the following important characterization of compatible linear systems.

**Theorem 4.46.** A linear system  $A \mathbf{x} = \mathbf{b}$  has a solution if and only if  $\mathbf{b}$  is orthogonal to the cokernel of  $A$ .

Indeed, the system has a solution if and only if the right-hand side belongs to the image of the coefficient matrix,  $\mathbf{b} \in \text{img } A$ , which, by (4.46), requires that  $\mathbf{b}$  be orthogonal to its cokernel. Thus, the compatibility conditions for the linear system  $A \mathbf{x} = \mathbf{b}$  can be written in the form

$$\mathbf{y} \cdot \mathbf{b} = 0 \quad \text{for every } \mathbf{y} \text{ satisfying} \quad A^T \mathbf{y} = \mathbf{0}. \quad (4.47)$$

In practice, one only needs to check orthogonality of  $\mathbf{b}$  with respect to a basis  $\mathbf{y}_1, \dots, \mathbf{y}_{m-r}$  of the cokernel, leading to a system of  $m-r$  compatibility constraints

$$\mathbf{y}_i \cdot \mathbf{b} = 0, \quad i = 1, \dots, m-r. \quad (4.48)$$

Here  $r = \text{rank } A$  denotes the rank of the coefficient matrix, and so  $m-r$  is also the number of all zero rows in the row echelon form of  $A$ . Hence, (4.48) contains precisely the same number of constraints as would be derived using Gaussian Elimination.

Theorem 4.46 is known as the *Fredholm alternative*, named after the Swedish mathematician Ivar Fredholm. His primary motivation was to solve linear integral equations, but his compatibility criterion was recognized to be a general property of linear systems, including linear algebraic systems, linear differential equations, linear boundary value problems, and so on.

**Example 4.47.** In Example 2.40, we analyzed the linear system  $A \mathbf{x} = \mathbf{b}$  with coefficient matrix  $A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \\ 1 & -2 & 3 \end{pmatrix}$ . Using direct Gaussian Elimination, we were led to a single compatibility condition, namely  $-b_1 + 2b_2 + b_3 = 0$ , required for the system to have a solution. We now understand the meaning behind this equation: it is telling us that the right-hand side  $\mathbf{b}$  must be orthogonal to the cokernel of  $A$ . The cokernel is determined by

solving the homogeneous adjoint system  $A^T \mathbf{y} = \mathbf{0}$ , and is the line spanned by the vector  $\mathbf{y}_1 = (-1, 2, 1)^T$ . Thus, the compatibility condition requires that  $\mathbf{b}$  be orthogonal to  $\mathbf{y}_1$ , in accordance with the Fredholm alternative (4.48).

**Example 4.48.** Let us determine the compatibility conditions for the linear system

$$x_1 - x_2 + 3x_3 = b_1, \quad -x_1 + 2x_2 - 4x_3 = b_2, \quad 2x_1 + 3x_2 + x_3 = b_3, \quad x_1 + 2x_3 = b_4,$$

by computing the cokernel of its coefficient matrix

$$A = \begin{pmatrix} 1 & -1 & 3 \\ -1 & 2 & -4 \\ 2 & 3 & 1 \\ 1 & 0 & 2 \end{pmatrix}.$$

We need to solve the homogeneous adjoint system  $A^T \mathbf{y} = \mathbf{0}$ , namely

$$y_1 - y_2 + 2y_3 + y_4 = 0, \quad -y_1 + 2y_2 + 3y_3 = 0, \quad 3y_1 - 4y_2 + y_3 + 2y_4 = 0.$$

Applying Gaussian Elimination, we deduce that the general solution

$$\mathbf{y} = y_3 (-7, -5, 1, 0)^T + y_4 (-2, -1, 0, 1)^T$$

is a linear combination (whose coefficients are the free variables) of the two basis vectors for  $\text{coker } A$ . Thus, the Fredholm compatibility conditions (4.48) are obtained by taking their dot products with the right-hand side of the original system:

$$-7b_1 - 5b_2 + b_3 = 0, \quad -2b_1 - b_2 + b_4 = 0.$$

The reader can check that these are indeed the same compatibility conditions that result from a direct Gaussian Elimination on the augmented matrix  $(A \mid \mathbf{b})$ .

**Remark.** Conversely, rather than solving the homogeneous adjoint system, we can use Gaussian Elimination on the augmented matrix  $(A \mid \mathbf{b})$  to determine the  $m - r$  basis vectors  $\mathbf{y}_1, \dots, \mathbf{y}_{m-r}$  for  $\text{coker } A$ . They are formed from the coefficients of  $b_1, \dots, b_m$  in the  $m - r$  consistency conditions  $\mathbf{y}_i \cdot \mathbf{b} = 0$  for  $i = 1, \dots, m - r$ , arising from the all zero rows in the reduced row echelon form.

We are now very close to a full understanding of the fascinating geometry that lurks behind the simple algebraic operation of multiplying a vector  $\mathbf{x} \in \mathbb{R}^n$  by an  $m \times n$  matrix, resulting in a vector  $\mathbf{b} = A\mathbf{x} \in \mathbb{R}^m$ . Since the kernel and coimage of  $A$  are orthogonal complements in the domain space  $\mathbb{R}^n$ , Proposition 4.41 tells us that we can uniquely decompose  $\mathbf{x} = \mathbf{w} + \mathbf{z}$ , where  $\mathbf{w} \in \text{coimg } A$ , while  $\mathbf{z} \in \ker A$ . Since  $A\mathbf{z} = \mathbf{0}$ , we have

$$\mathbf{b} = A\mathbf{x} = A(\mathbf{w} + \mathbf{z}) = A\mathbf{w}.$$

Therefore, we can regard multiplication by  $A$  as a combination of two operations:

- (i) The first is an orthogonal projection onto the coimage of  $A$  taking  $\mathbf{x}$  to  $\mathbf{w}$ .
- (ii) The second maps a vector in  $\text{coimg } A \subset \mathbb{R}^n$  to a vector in  $\text{img } A \subset \mathbb{R}^m$ , taking the orthogonal projection  $\mathbf{w}$  to the image vector  $\mathbf{b} = A\mathbf{w} = A\mathbf{x}$ .

Moreover, if  $A$  has rank  $r$ , then both  $\text{img } A$  and  $\text{coimg } A$  are  $r$ -dimensional subspaces, albeit of different vector spaces. Each vector  $\mathbf{b} \in \text{img } A$  corresponds to a *unique* vector  $\mathbf{w} \in \text{coimg } A$ . Indeed, if  $\mathbf{w}, \tilde{\mathbf{w}} \in \text{coimg } A$  satisfy  $\mathbf{b} = A\mathbf{w} = A\tilde{\mathbf{w}}$ , then  $A(\mathbf{w} - \tilde{\mathbf{w}}) = \mathbf{0}$ , and hence  $\mathbf{w} - \tilde{\mathbf{w}} \in \ker A$ . But, since the kernel and the coimage are orthogonal complements,

the only vector that belongs to both is the zero vector, and hence  $\mathbf{w} = \tilde{\mathbf{w}}$ . In this manner, we have proved the first part of the following result; the second is left as Exercise 4.4.38.

**Theorem 4.49.** Multiplication by an  $m \times n$  matrix  $A$  of rank  $r$  defines a one-to-one correspondence between the  $r$ -dimensional subspaces  $\text{coimg } A \subset \mathbb{R}^n$  and  $\text{img } A \subset \mathbb{R}^m$ . Moreover, if  $\mathbf{v}_1, \dots, \mathbf{v}_r$  forms a basis of  $\text{coimg } A$  then their images  $A\mathbf{v}_1, \dots, A\mathbf{v}_r$  form a basis for  $\text{img } A$ .

In summary, the linear system  $A\mathbf{x} = \mathbf{b}$  has a solution if and only if  $\mathbf{b} \in \text{img } A$ , or, equivalently, is orthogonal to every vector  $\mathbf{y} \in \text{coker } A$ . If the compatibility conditions hold, then the system has a *unique* solution  $\mathbf{w} \in \text{coimg } A$  that, by the definition of the coimage, is a linear combination of the *rows* of  $A$ . The general solution to the system is  $\mathbf{x} = \mathbf{w} + \mathbf{z}$ , where  $\mathbf{w}$  is the particular solution belonging to the coimage, while  $\mathbf{z} \in \ker A$  is an arbitrary element of the kernel.

**Theorem 4.50.** A compatible linear system  $A\mathbf{x} = \mathbf{b}$  with  $\mathbf{b} \in \text{img } A = (\text{coker } A)^\perp$  has a unique solution  $\mathbf{w} \in \text{coimg } A$  satisfying  $A\mathbf{w} = \mathbf{b}$ . The general solution is  $\mathbf{x} = \mathbf{w} + \mathbf{z}$ , where  $\mathbf{z} \in \ker A$ . The particular solution  $\mathbf{w} \in \text{coimg } A$  is distinguished by the fact that it has the smallest Euclidean norm of all possible solutions:  $\|\mathbf{w}\| \leq \|\mathbf{x}\|$  whenever  $A\mathbf{x} = \mathbf{b}$ .

*Proof:* We have already established all but the last statement. Since the coimage and kernel are orthogonal subspaces, the norm of a general solution  $\mathbf{x} = \mathbf{w} + \mathbf{z}$  is

$$\|\mathbf{x}\|^2 = \|\mathbf{w} + \mathbf{z}\|^2 = \|\mathbf{w}\|^2 + 2\mathbf{w} \cdot \mathbf{z} + \|\mathbf{z}\|^2 = \|\mathbf{w}\|^2 + \|\mathbf{z}\|^2 \geq \|\mathbf{w}\|^2,$$

with equality if and only if  $\mathbf{z} = \mathbf{0}$ .

*Q.E.D.*

In practice, to determine the unique minimum-norm solution to a compatible linear system, we invoke the orthogonality of the coimage and kernel of the coefficient matrix. Thus, if  $\mathbf{z}_1, \dots, \mathbf{z}_{n-r}$  form a basis for  $\ker A$ , then the minimum-norm solution  $\mathbf{x} = \mathbf{w} \in \text{coimg } A$  is obtained by solving the enlarged system

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{z}_1^T \mathbf{x} = 0, \quad \dots \quad \mathbf{z}_{n-r}^T \mathbf{x} = 0. \quad (4.49)$$

The associated  $(m + n - r) \times n$  coefficient matrix is simply obtained by appending the (transposed) kernel vectors to the original matrix  $A$ . The resulting matrix is guaranteed to have maximum rank  $n$ , and so, assuming  $\mathbf{b} \in \text{img } A$ , the enlarged system has a unique solution, which is the minimum-norm solution to the original system  $A\mathbf{x} = \mathbf{b}$ .

**Example 4.51.** Consider the linear system

$$\begin{pmatrix} 1 & -1 & 2 & -2 \\ 0 & 1 & -2 & 1 \\ 1 & 3 & -5 & 2 \\ 5 & -1 & 9 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 4 \\ 6 \end{pmatrix}. \quad (4.50)$$

Applying the usual Gaussian Elimination algorithm, we discover that the coefficient matrix has rank 3, and its kernel is spanned by the single vector  $\mathbf{z}_1 = (1, -1, 0, 1)^T$ . The system itself is compatible; indeed, the right-hand side is orthogonal to the basis cokernel vector  $(2, 24, -7, 1)^T$ , and so satisfies the Fredholm condition (4.48). The general solution to the linear system is  $\mathbf{x} = (t, 3 - t, 1, t)^T$ , where  $t = w$  is the free variable.

To find the solution of minimum Euclidean norm, we can apply the algorithm described in the previous paragraph.<sup>†</sup> Thus, we supplement the original system by the constraint

$$(1 \ -1 \ 0 \ 1) \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = x - y + w = 0 \quad (4.51)$$

that the solution be orthogonal to the kernel basis vector. Solving the combined linear system (4.50–51) leads to the unique solution  $\mathbf{x} = \mathbf{w} = (1, 2, 1, 1)^T$ , obtained by setting the free variable  $t$  equal to 1. Let us check that its norm is indeed the smallest among all solutions to the original system:

$$\|\mathbf{w}\| = \sqrt{7} \leq \|\mathbf{x}\| = \|(t, 3-t, 1, t)^T\| = \sqrt{3t^2 - 6t + 10},$$

where the quadratic function inside the square root achieves its minimum value of  $\sqrt{7}$  at  $t = 1$ . It is further distinguished as the only solution that can be expressed as a linear combination of the rows of the coefficient matrix:

$$\begin{aligned} \mathbf{w}^T &= (1, 2, 1, 1) \\ &= -4(1, -1, 2, -2) - 17(0, 1, -2, 1) + 5(1, 3, -5, 2), \end{aligned}$$

meaning that  $\mathbf{w}$  lies in the coimage of the coefficient matrix.

## Exercises

4.4.29. For each of the following matrices  $A$ , (i) find a basis for each of the four fundamental subspaces; (ii) verify that the image and cokernel are orthogonal complements; (iii) verify that the coimage and kernel are orthogonal complements:

$$\begin{aligned} (a) \quad & \begin{pmatrix} 1 & -2 \\ 2 & -4 \end{pmatrix}, \quad (b) \quad \begin{pmatrix} 5 & 0 \\ 1 & 2 \\ 0 & 2 \end{pmatrix}, \quad (c) \quad \begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & -3 \\ -2 & 3 & 0 \end{pmatrix}, \quad (d) \quad \begin{pmatrix} 1 & 2 & 0 & 1 \\ -1 & 1 & 3 & 1 \\ 0 & 3 & 3 & 2 \end{pmatrix}, \\ (e) \quad & \begin{pmatrix} 3 & 1 & 4 & 2 & 7 \\ 1 & 1 & 2 & 0 & 3 \\ 5 & 2 & 7 & 3 & 12 \end{pmatrix}, \quad (f) \quad \begin{pmatrix} 1 & 3 & 0 & -2 \\ -2 & 1 & 2 & 3 \\ -3 & 5 & 4 & 4 \\ 1 & -4 & -2 & -1 \end{pmatrix}, \quad (g) \quad \begin{pmatrix} -1 & 2 & 2 & -1 \\ 2 & -4 & -5 & 2 \\ -3 & 6 & 2 & -3 \\ 1 & -2 & -3 & 1 \\ -2 & 4 & -5 & -2 \end{pmatrix}. \end{aligned}$$

4.4.30. For each of the following matrices, use Gaussian elimination on the augmented matrix  $(A \mid \mathbf{b})$  to determine a basis for its cokernel:

$$(a) \quad \begin{pmatrix} 9 & -6 \\ 6 & -4 \end{pmatrix}, \quad (b) \quad \begin{pmatrix} 1 & 3 \\ 2 & 6 \\ -3 & -9 \end{pmatrix}, \quad (c) \quad \begin{pmatrix} 1 & 1 & 3 \\ -1 & 1 & -2 \\ -1 & 3 & 6 \end{pmatrix}, \quad (d) \quad \begin{pmatrix} 1 & -2 & -2 \\ 0 & -1 & 3 \\ 2 & -5 & -1 \\ -2 & 2 & 10 \end{pmatrix}.$$

4.4.31. Let  $A = \begin{pmatrix} 1 & -2 & 2 & -1 \\ -2 & 4 & -3 & 5 \\ -1 & 2 & 0 & 7 \end{pmatrix}$ . (a) Find a basis for  $\text{coimg } A$ . (b) Use Theorem 4.49

to find a basis of  $\text{img } A$ . (c) Write each column of  $A$  as a linear combination of the basis vectors you found in part (b).

<sup>†</sup> An alternative is to orthogonally project the general solution onto the coimage. The result is the same.

- 4.4.32. Write down the compatibility conditions on the following systems of linear equations by first computing a basis for the cokernel of the coefficient matrix. (a)  $2x + y = a$ ,  $x + 4y = b$ ,  $-3x + 2y = c$ ; (b)  $x + 2y + 3z = a$ ,  $-x + 5y - 2z = b$ ,  $2x - 3y + 5z = c$ ; (c)  $x_1 + 2x_2 + 3x_3 = b_1$ ,  $x_2 + 2x_3 = b_2$ ,  $3x_1 + 5x_2 + 7x_3 = b_3$ ,  $-2x_1 + x_2 + 4x_3 = b_4$ ; (d)  $x - 3y + 2z + w = a$ ,  $4x - 2y + 2z + 3w = b$ ,  $5x - 5y + 4z + 4w = c$ ,  $2x + 4y - 2z + w = d$ .

- 4.4.33. For each of the following  $m \times n$  matrices, decompose the first standard basis vector  $\mathbf{e}_1 = \mathbf{w} + \mathbf{z} \in \mathbb{R}^n$ , where  $\mathbf{w} \in \text{coimg } A$  and  $\mathbf{z} \in \ker A$ . Verify your answer by expressing  $\mathbf{w}$  as a linear combination of the rows of  $A$ .

$$(a) \begin{pmatrix} 1 & -2 & 1 \\ 2 & -3 & 2 \end{pmatrix}, (b) \begin{pmatrix} 1 & 1 & 2 \\ -1 & 0 & -1 \\ -2 & -1 & -3 \end{pmatrix}, (c) \begin{pmatrix} 1 & -1 & 0 & 3 \\ 2 & 1 & 3 & 3 \\ 1 & 2 & 3 & 0 \end{pmatrix}, (d) \begin{pmatrix} -1 & 1 & 1 & -1 & 2 \\ -3 & 2 & -1 & -2 & 0 \end{pmatrix}.$$

- 4.4.34. For each of the following linear systems, (i) verify compatibility using the Fredholm alternative, (ii) find the general solution, and (iii) find the solution of minimum Euclidean norm:

$$(a) \begin{array}{l} 2x - 4y = -6, \\ -x + 2y = 3, \end{array} \quad (b) \begin{array}{l} 2x + 3y = -1, \\ 3x + 7y = 1, \\ -3x + 2y = 8, \end{array} \quad (c) \begin{array}{l} 6x - 3y + 9z = 12, \\ 2x - y + 3z = 4, \end{array}$$

$$(d) \begin{array}{l} x + 3y + 5z = 3, \\ -x + 4y + 9z = 11, \\ 2x + 3y + 4z = 0, \end{array} \quad (e) \begin{array}{l} x_1 - 3x_2 + 7x_3 = -8, \\ 2x_1 + x_2 = 5, \\ 4x_1 - 3x_2 + 10x_3 = -5, \\ -2x_1 + 2x_2 - 6x_3 = 4. \end{array} \quad (f) \begin{array}{l} x - y + 2z + 3w = 5, \\ 3x - 3y + 5z + 7w = 13, \\ -2x + 2y + z + 4w = 0. \end{array}$$

- 4.4.35. Show that if  $A = A^T$  is a symmetric matrix, then  $A\mathbf{x} = \mathbf{b}$  has a solution if and only if  $\mathbf{b}$  is orthogonal to  $\ker A$ .

- ◇ 4.4.36. Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n$  span a subspace  $V \subset \mathbb{R}^m$ . Prove that  $\mathbf{w}$  is orthogonal to  $V$  if and only if  $\mathbf{w} \in \text{coker } A$ , where  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  is the matrix with the indicated columns.

- 4.4.37. Let  $A = \begin{pmatrix} 1 & -1 & 0 & 2 \\ 2 & -2 & 0 & 4 \\ -1 & 1 & 1 & -1 \\ 0 & 0 & 2 & 2 \end{pmatrix}$ . (a) Find an orthogonal basis for  $\text{coimg } A$ . (b) Find an orthogonal basis for  $\ker A$ . (c) If you combine your bases from parts (a) and (b), do you get an orthogonal basis of  $\mathbb{R}^4$ ? Why or why not?

- ◇ 4.4.38. Prove that if  $\mathbf{v}_1, \dots, \mathbf{v}_r$  are a basis of  $\text{coimg } A$ , then their images  $A\mathbf{v}_1, \dots, A\mathbf{v}_r$  are a basis for  $\text{img } A$ .

- 4.4.39. *True or false:* The standard algorithm for finding a basis for  $\ker A$  will always produce an orthogonal basis.

- ◇ 4.4.40. Is Theorem 4.45 true as stated for complex matrices? If not, can you formulate a similar theorem that is true? What is the Fredholm alternative for complex matrices?

## 4.5 Orthogonal Polynomials

Orthogonal and orthonormal bases play, if anything, an even more essential role in function spaces. Unlike the Euclidean space  $\mathbb{R}^n$ , most of the obvious bases of a typical (finite-dimensional) function space are not orthogonal with respect to any natural inner product. Thus, the computation of an orthonormal basis of functions is a critical step towards simplification of the analysis. The Gram–Schmidt algorithm, in any of the above formulations, can be successfully applied to construct suitably orthogonal functions. The most impor-

tant examples are the classical orthogonal polynomials that arise in approximation and interpolation theory. Other orthogonal systems of functions play starring roles in Fourier analysis and its generalizations, including wavelets, in quantum mechanics, in the solution of partial differential equations by separation of variables, and a host of further applications in mathematics, physics, engineering, numerical analysis, etc., [43, 54, 62, 61, 77, 79, 88].

### The Legendre Polynomials

We shall construct an orthonormal basis for the vector space  $\mathcal{P}^{(n)}$  of polynomials of degree  $\leq n$ . For definiteness, the construction will be based on the  $L^2$  inner product

$$\langle p, q \rangle = \int_{-1}^1 p(t) q(t) dt \quad (4.52)$$

on the interval  $[-1, 1]$ . The underlying method will work on any other bounded interval, as well as for weighted inner products, but (4.52) is of particular importance. We shall apply the Gram–Schmidt orthogonalization process to the elementary, but non-orthogonal monomial basis  $1, t, t^2, \dots, t^n$ . Because

$$\langle t^k, t^l \rangle = \int_{-1}^1 t^{k+l} dt = \begin{cases} \frac{2}{k+l+1}, & k+l \text{ even,} \\ 0, & k+l \text{ odd,} \end{cases} \quad (4.53)$$

odd-degree monomials are orthogonal to those of even degree, but that is all. We will use  $q_0(t), q_1(t), \dots, q_n(t)$  to denote the resulting orthogonal polynomials. We begin by setting

$$q_0(t) = 1, \quad \text{with} \quad \|q_0\|^2 = \int_{-1}^1 q_0(t)^2 dt = 2.$$

According to formula (4.17), the next orthogonal basis polynomial is

$$q_1(t) = t - \frac{\langle t, q_0 \rangle}{\|q_0\|^2} q_0(t) = t, \quad \text{with} \quad \|q_1\|^2 = \frac{2}{3}.$$

In general, the Gram–Schmidt formula (4.19) says we should define

$$q_k(t) = t^k - \sum_{j=0}^{k-1} \frac{\langle t^k, q_j \rangle}{\|q_j\|^2} q_j(t) \quad \text{for} \quad k = 1, 2, \dots.$$

We can thus recursively compute the next few orthogonal polynomials:

$$\begin{aligned} q_2(t) &= t^2 - \frac{1}{3}, & \|q_2\|^2 &= \frac{8}{45}, \\ q_3(t) &= t^3 - \frac{3}{5}t, & \|q_3\|^2 &= \frac{8}{175}, \\ q_4(t) &= t^4 - \frac{6}{7}t^2 + \frac{3}{35}, & \|q_4\|^2 &= \frac{128}{11025}, \\ q_5(t) &= t^5 - \frac{10}{9}t^3 + \frac{5}{21}t & \|q_5\|^2 &= \frac{128}{43659}, \end{aligned} \quad (4.54)$$

and so on. The reader can verify that they satisfy the orthogonality conditions

$$\langle q_i, q_j \rangle = \int_{-1}^1 q_i(t) q_j(t) dt = 0, \quad i \neq j.$$

The resulting polynomials  $q_0, q_1, q_2, \dots$  are known as the *monic<sup>†</sup> Legendre polynomials*, in honor of the eighteenth-century French mathematician Adrien-Marie Legendre, who first

---

<sup>†</sup> A polynomial is called *monic* if its leading coefficient is equal to 1.

used them for studying Newtonian gravitation. Since the first  $n$  Legendre polynomials, namely  $q_0, \dots, q_{n-1}$  span the subspace  $\mathcal{P}^{(n-1)}$  of polynomials of degree  $\leq n-1$ , the next one,  $q_n$ , can be characterized as the unique monic polynomial that is orthogonal to every polynomial of degree  $\leq n-1$ :

$$\langle t^k, q_n \rangle = 0, \quad k = 0, \dots, n-1. \quad (4.55)$$

Since the monic Legendre polynomials form a basis for the space of polynomials, we can uniquely rewrite any polynomial of degree  $n$  as a linear combination:

$$p(t) = c_0 q_0(t) + c_1 q_1(t) + \dots + c_n q_n(t). \quad (4.56)$$

In view of the general orthogonality formula (4.7), the coefficients are simply given by inner products

$$c_k = \frac{\langle p, q_k \rangle}{\|q_k\|^2} = \frac{1}{\|q_k\|^2} \int_{-1}^1 p(t) q_k(t) dt, \quad k = 0, \dots, n. \quad (4.57)$$

For example,

$$t^4 = q_4(t) + \frac{6}{7} q_2(t) + \frac{1}{5} q_0(t) = \left( t^4 - \frac{6}{7} t^2 + \frac{3}{35} \right) + \frac{6}{7} \left( t^2 - \frac{1}{3} \right) + \frac{1}{5},$$

where the coefficients can be obtained either directly or via (4.57):

$$c_4 = \frac{11025}{128} \int_{-1}^1 t^4 q_4(t) dt = 1, \quad c_3 = \frac{175}{8} \int_{-1}^1 t^4 q_3(t) dt = 0, \quad \text{and so on.}$$

The classical *Legendre polynomials*, [59], are certain scalar multiples, namely

$$P_k(t) = \frac{(2k)!}{2^k (k!)^2} q_k(t), \quad k = 0, 1, 2, \dots, \quad (4.58)$$

and so also define a system of orthogonal polynomials. The multiple is fixed by the requirement that

$$P_k(1) = 1, \quad (4.59)$$

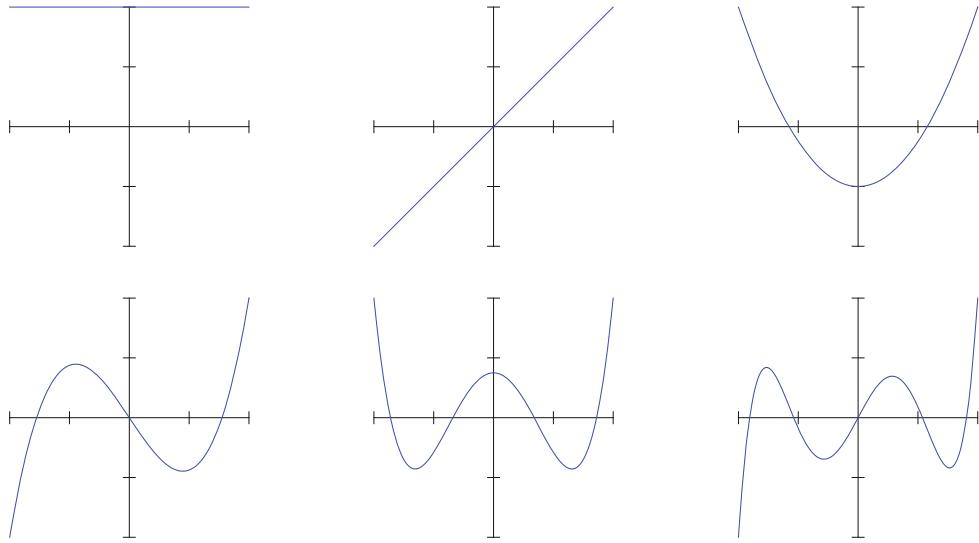
which is not so important here, but does play a role in other applications. The first few classical Legendre polynomials are

$$\begin{aligned} P_0(t) &= 1, & \|P_0\|^2 &= 2, \\ P_1(t) &= t, & \|P_1\|^2 &= \frac{2}{3}, \\ P_2(t) &= \frac{3}{2}t^2 - \frac{1}{2}, & \|P_2\|^2 &= \frac{2}{5}, \\ P_3(t) &= \frac{5}{2}t^3 - \frac{3}{2}t, & \|P_3\|^2 &= \frac{2}{7}, \\ P_4(t) &= \frac{35}{8}t^4 - \frac{15}{4}t^2 + \frac{3}{8}, & \|P_4\|^2 &= \frac{2}{9}, \\ P_5(t) &= \frac{63}{8}t^5 - \frac{35}{4}t^3 + \frac{15}{8}t, & \|P_5\|^2 &= \frac{2}{11}, \end{aligned} \quad (4.60)$$

and are graphed in Figure 4.8. There is, in fact, an explicit formula for the Legendre polynomials, due to the early nineteenth-century mathematician, banker, and social reformer Olinde Rodrigues.

**Theorem 4.52.** The *Rodrigues formula* for the classical Legendre polynomials is

$$P_k(t) = \frac{1}{2^k k!} \frac{d^k}{dt^k} (t^2 - 1)^k, \quad \|P_k\| = \sqrt{\frac{2}{2k+1}}, \quad k = 0, 1, 2, \dots. \quad (4.61)$$



**Figure 4.8.** The Legendre Polynomials  $P_0(t), \dots, P_5(t)$ .

Thus, for example,

$$P_4(t) = \frac{1}{16 \cdot 4!} \frac{d^4}{dt^4} (t^2 - 1)^4 = \frac{1}{384} \frac{d^4}{dt^4} (t^2 - 1)^4 = \frac{35}{8} t^4 - \frac{15}{4} t^2 + \frac{3}{8}.$$

*Proof of Theorem 4.52:* Let

$$R_{j,k}(t) = \frac{d^j}{dt^j} (t^2 - 1)^k, \quad (4.62)$$

which is evidently a polynomial of degree  $2k - j$ . In particular, the Rodrigues formula (4.61) claims that  $P_k(t)$  is a multiple of  $R_{k,k}(t)$ . Note that

$$\frac{d}{dt} R_{j,k}(t) = R_{j+1,k}(t). \quad (4.63)$$

Moreover,

$$R_{j,k}(1) = 0 = R_{j,k}(-1) \quad \text{whenever} \quad j < k, \quad (4.64)$$

since, by the product rule, differentiating  $(t^2 - 1)^k$  a total of  $j < k$  times still leaves at least one factor of  $t^2 - 1$  in each summand, which therefore vanishes at  $t = \pm 1$ . In order to complete the proof of the first formula, let us establish the following result:

**Lemma 4.53.** If  $j \leq k$ , then the polynomial  $R_{j,k}(t)$  is orthogonal to all polynomials of degree  $\leq j - 1$ .

*Proof:* In other words,

$$\langle t^i, R_{j,k} \rangle = \int_{-1}^1 t^i R_{j,k}(t) dt = 0, \quad \text{for all} \quad 0 \leq i < j \leq k. \quad (4.65)$$

Since  $j > 0$ , we use (4.63) to write  $R_{j,k}(t) = R'_{j-1,k}(t)$ . Integrating by parts,

$$\begin{aligned}\langle t^i, R_{j,k} \rangle &= \int_{-1}^1 t^i R'_{j-1,k}(t) dt \\ &= i t^i R_{j-1,k}(t) \Big|_{t=-1}^1 - i \int_{-1}^1 t^{i-1} R_{j-1,k}(t) dt = -i \langle t^{i-1}, R_{j-1,k} \rangle,\end{aligned}$$

where the boundary terms vanish owing to (4.64). In particular, setting  $i = 0$  proves  $\langle 1, R_{j,k} \rangle = 0$  for all  $j > 0$ . We then repeat the process, and, eventually, for any  $j > i$ ,

$$\begin{aligned}\langle t^i, R_{j,k} \rangle &= -i \langle t^{i-1}, R_{j-1,k} \rangle \\ &= i(i-1) \langle t^{i-2}, R_{j-2,k} \rangle = \dots = (-1)^i i(i-1) \dots 3 \cdot 2 \langle 1, R_{j-i,k} \rangle = 0,\end{aligned}$$

completing the proof. *Q.E.D.*

In particular,  $R_{k,k}(t)$  is a polynomial of degree  $k$  that is orthogonal to every polynomial of degree  $\leq k-1$ . By our earlier remarks, this implies that it must be a constant multiple,

$$R_{k,k}(t) = c_k P_k(t),$$

of the  $k^{\text{th}}$  Legendre polynomial. To determine  $c_k$ , we need only compare the leading terms:

$$R_{k,k}(t) = \frac{d^k}{dt^k} (t^2 - 1)^k = \frac{d^k}{dt^k} (t^{2k} + \dots) = \frac{(2k)!}{k!} t^k + \dots, \text{ while } P_k(t) = \frac{(2k)!}{2^k (k!)^2} t^{2k} + \dots.$$

We conclude that  $c_k = 2^k k!$ , which proves the first formula in (4.61). The proof of the formula for  $\|P_k\|$  can be found in Exercise 4.5.9. *Q.E.D.*

The Legendre polynomials play an important role in many aspects of applied mathematics, including numerical analysis, least squares approximation of functions, and the solution of partial differential equations, [61].

## Exercises

4.5.1. Write the following polynomials as linear combinations of monic Legendre polynomials.

Use orthogonality to compute the coefficients: (a)  $t^3$ , (b)  $t^4 + t^2$ , (c)  $7t^4 + 2t^3 - t$ .

4.5.2. (a) Find the monic Legendre polynomial of degree 5 using the Gram–Schmidt process.

Check your answer using the Rodrigues formula. (b) Use orthogonality to write  $t^5$  as a linear combination of Legendre polynomials. (c) Repeat the exercise for degree 6.

◇ 4.5.3. (a) Explain why  $q_n$  is the unique monic polynomial that satisfies (4.55). (b) Use this characterization to directly construct  $q_5(t)$ .

4.5.4. Prove that the even (odd) degree Legendre polynomials are even (odd) functions of  $t$ .

4.5.5. Prove that if  $p(t) = p(-t)$  is an even polynomial, then all the odd-order coefficients  $c_{2j+1} = 0$  in its Legendre expansion (4.56) vanish.

4.5.6. Write out an explicit Rodrigues-type formula for the monic Legendre polynomial  $q_k(t)$  and its norm.

4.5.7. Write out an explicit Rodrigues-type formula for an *orthonormal basis*  $Q_0(t), \dots, Q_n(t)$  for the space of polynomials of degree  $\leq n$  under the inner product (4.52).

◇ 4.5.8. Use the Rodrigues formula to prove (4.59). *Hint:* Write  $(t^2 - 1)^k = (t - 1)^k (t + 1)^k$ .

- ◇ 4.5.9. A proof of the formula in (4.61) for the norms of the Legendre polynomials is based on the following steps. (a) First, prove that  $\|R_{k,k}\|^2 = (-1)^k (2k)! \int_{-1}^1 (t^2 - 1)^k dt$  by a repeated integration by parts. (b) Second, prove that  $\int_{-1}^1 (t^2 - 1)^k dt = (-1)^k \frac{2^{2k+1} (k!)^2}{(2k+1)!}$  by using the change of variables  $t = \cos \theta$  in the integral. The resulting trigonometric integral can be done by another repeated integration by parts. (c) Finally, use the Rodrigues formula to complete the proof.
- ◇ 4.5.10. (a) Find the roots,  $P_n(t) = 0$ , of the Legendre polynomials  $P_2, P_3$  and  $P_4$ . (b) Prove that for  $0 \leq j \leq k$ , the polynomial  $R_{j,k}(t)$  defined in (4.62) has roots of order  $k - j$  at  $t = \pm 1$ , and  $j$  additional simple roots lying between  $-1$  and  $1$ . Hint: Use induction on  $j$  and Rolle's Theorem from calculus, [2, 78]. (c) Conclude that all  $k$  roots of the Legendre polynomial  $P_k(t)$  are real and simple, and that they lie in the interval  $-1 < t < 1$ .

## Other Systems of Orthogonal Polynomials

The standard Legendre polynomials form an orthogonal system with respect to the  $L^2$  inner product on the interval  $[-1, 1]$ . Dealing with any other interval, or, more generally, a weighted inner product, leads to a different, suitably adapted collection of orthogonal polynomials. In all cases, applying the Gram–Schmidt process to the standard monomials  $1, t, t^2, t^3, \dots$  will produce the desired orthogonal system.

**Example 4.54.** In this example, we construct orthogonal polynomials for the weighted inner product<sup>†</sup>

$$\langle f, g \rangle = \int_0^\infty f(t) g(t) e^{-t} dt \quad (4.66)$$

on the interval  $[0, \infty)$ . A straightforward integration by parts proves that

$$\int_0^\infty t^k e^{-t} dt = k!, \quad \text{and hence} \quad \langle t^i, t^j \rangle = (i+j)!, \quad \|t^i\|^2 = (2i)!. \quad (4.67)$$

We apply the Gram–Schmidt process to construct a system of orthogonal polynomials for this inner product. The first few are

$$\begin{aligned} q_0(t) &= 1, & \|q_0\|^2 &= 1, \\ q_1(t) &= t - \frac{\langle t, q_0 \rangle}{\|q_0\|^2} q_0(t) = t - 1, & \|q_1\|^2 &= 1, \\ q_2(t) &= t^2 - \frac{\langle t^2, q_0 \rangle}{\|q_0\|^2} q_0(t) - \frac{\langle t^2, q_1 \rangle}{\|q_1\|^2} q_1(t) = t^2 - 4t + 2, & \|q_2\|^2 &= 4, \\ q_3(t) &= t^3 - 9t^2 + 18t - 6, & \|q_3\|^2 &= 36. \end{aligned} \quad (4.68)$$

The resulting orthogonal polynomials are known as the (monic) *Laguerre polynomials*, named after the nineteenth-century French mathematician Edmond Laguerre, [59].

<sup>†</sup> The functions  $f, g$  must not grow too rapidly as  $t \rightarrow \infty$  in order that the inner product be defined. For example, polynomial growth, meaning  $|f(t)|, |g(t)| \leq Ct^N$  for  $t \gg 0$  and some  $C > 0, 0 \leq N < \infty$ , suffices.

In some cases, a change of variables may be used to relate systems of orthogonal polynomials and thereby circumvent the Gram–Schmidt computation. Suppose, for instance, that our goal is to construct an orthogonal system of polynomials for the  $L^2$  inner product

$$\langle\langle f, g \rangle\rangle = \int_a^b f(t) g(t) dt$$

on the interval  $[a, b]$ . The key remark is that we can map the interval  $[-1, 1]$  to  $[a, b]$  by a simple change of variables of the form  $s = \alpha + \beta t$ . Specifically,

$$s = \frac{2t - b - a}{b - a} \quad \text{will change} \quad a \leq t \leq b \quad \text{to} \quad -1 \leq s \leq 1. \quad (4.69)$$

It therefore changes functions  $F(s), G(s)$ , defined for  $-1 \leq s \leq 1$ , into functions

$$f(t) = F\left(\frac{2t - b - a}{b - a}\right), \quad g(t) = G\left(\frac{2t - b - a}{b - a}\right), \quad (4.70)$$

defined for  $a \leq t \leq b$ . Moreover, when integrating, we have  $ds = \frac{2}{b - a} dt$ , and so the inner products are related by

$$\begin{aligned} \langle f, g \rangle &= \int_a^b f(t) g(t) dt = \int_a^b F\left(\frac{2t - b - a}{b - a}\right) G\left(\frac{2t - b - a}{b - a}\right) dt \\ &= \int_{-1}^1 F(s) G(s) \frac{b - a}{2} ds = \frac{b - a}{2} \langle F, G \rangle, \end{aligned} \quad (4.71)$$

where the final  $L^2$  inner product is over the interval  $[-1, 1]$ . In particular, the change of variables maintains orthogonality, while rescaling the norms; explicitly,

$$\langle f, g \rangle = 0 \quad \text{if and only if} \quad \langle F, G \rangle = 0, \quad \text{while} \quad \|f\| = \sqrt{\frac{b - a}{2}} \|F\|. \quad (4.72)$$

Moreover, if  $F(s)$  is a polynomial of degree  $n$  in  $s$ , then  $f(t)$  is a polynomial of degree  $n$  in  $t$  and conversely. Let us apply these observations to the Legendre polynomials:

**Proposition 4.55.** The transformed Legendre polynomials

$$\tilde{P}_k(t) = P_k\left(\frac{2t - b - a}{b - a}\right), \quad \|\tilde{P}_k\| = \sqrt{\frac{b - a}{2k + 1}}, \quad k = 0, 1, 2, \dots, \quad (4.73)$$

form an orthogonal system of polynomials with respect to the  $L^2$  inner product on the interval  $[a, b]$ .

**Example 4.56.** Consider the  $L^2$  inner product  $\langle\langle f, g \rangle\rangle = \int_0^1 f(t) g(t) dt$ . The map  $s = 2t - 1$  will change  $0 \leq t \leq 1$  to  $-1 \leq s \leq 1$ . According to Proposition 4.55, this change of variables will convert the Legendre polynomials  $P_k(s)$  into an orthogonal system of polynomials on  $[0, 1]$ , namely

$$\tilde{P}_k(t) = P_k(2t - 1), \quad \text{with corresponding } L^2 \text{ norms} \quad \|\tilde{P}_k\| = \sqrt{\frac{1}{2k + 1}}.$$

The first few are

$$\begin{aligned} \tilde{P}_0(t) &= 1, & \tilde{P}_3(t) &= 20t^3 - 30t^2 + 12t - 1, \\ \tilde{P}_1(t) &= 2t - 1, & \tilde{P}_4(t) &= 70t^4 - 140t^3 + 90t^2 - 20t + 1, \\ \tilde{P}_2(t) &= 6t^2 - 6t + 1, & \tilde{P}_5(t) &= 252t^5 - 630t^4 + 560t^3 - 210t^2 + 30t - 1. \end{aligned} \quad (4.74)$$

Alternatively, one can derive these formulas through a direct application of the Gram–Schmidt process.

## Exercises

4.5.11. Construct polynomials  $P_0, P_1, P_2$ , and  $P_3$  of degree 0, 1, 2, and 3, respectively, that are orthogonal with respect to the inner products (a)  $\langle f, g \rangle = \int_1^2 f(t) g(t) dt$ , (b)  $\langle f, g \rangle = \int_0^1 f(t) g(t) t dt$ , (c)  $\langle f, g \rangle = \int_{-1}^1 f(t) g(t) t^2 dt$ , (d)  $\langle f, g \rangle = \int_{-\infty}^{\infty} f(t) g(t) e^{-|t|} dt$ .

4.5.12. Find the first four orthogonal polynomials on the interval  $[0, 1]$  for the weighted  $L^2$  inner product with weight  $w(t) = t^2$ .

4.5.13. Write down an orthogonal basis for vector space  $\mathcal{P}^{(5)}$  of quintic polynomials under the inner product  $\langle f, g \rangle = \int_{-2}^2 f(t) g(t) dt$ .

4.5.14. Use the Gram–Schmidt process based on the  $L^2$  inner product on  $[0, 1]$  to construct a system of orthogonal polynomials of degree  $\leq 4$ . Verify that your polynomials are multiples of the modified Legendre polynomials found in Example 4.56.

4.5.15. Find the first four orthogonal polynomials under the Sobolev  $H^1$  inner product

$$\langle f, g \rangle = \int_{-1}^1 [f(t) g(t) + f'(t) g'(t)] dt; \text{ cf. Exercise 3.1.27.}$$

◇ 4.5.16. Prove the formula for  $\|\tilde{P}_k\|$  in (4.73).

4.5.17. Find the monic Laguerre polynomials of degrees 4 and 5 and their norms.

◇ 4.5.18. Prove the integration formula (4.67).

◇ 4.5.19. (a) The physicists' *Hermite polynomials* are orthogonal with respect to the inner product  $\langle f, g \rangle = \int_{-\infty}^{\infty} f(t) g(t) e^{-t^2} dt$ . Find the first five monic Hermite polynomials.

*Hint:*  $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$ . (b) The probabilists prefer to use the inner product

$\langle f, g \rangle = \int_{-\infty}^{\infty} f(t) g(t) e^{-t^2/2} dt$ . Find the first five of their monic Hermite polynomials.

(c) Can you find a change of variables that transforms the physicists' versions to the probabilists' versions?

◇ 4.5.20. The *Chebyshev polynomials*: (a) Prove that  $T_n(t) = \cos(n \arccos t)$ ,  $n = 0, 1, 2, \dots$ , form a system of orthogonal polynomials under the weighted inner product

$$\langle f, g \rangle = \int_{-1}^1 \frac{f(t) g(t) dt}{\sqrt{1 - t^2}}. \quad (4.75)$$

(b) What is  $\|T_n\|$ ? (c) Write out the formulas for  $T_0(t), \dots, T_6(t)$  and plot their graphs.

4.5.21. Does the Gram–Schmidt process for the inner product (4.75) lead to the Chebyshev polynomials  $T_n(t)$  defined in the preceding exercise? Explain why or why not.

4.5.22. Find two functions that form an orthogonal basis for the space of the solutions to the differential equation  $y'' - 3y' + 2y = 0$  under the  $L^2$  inner product on  $[0, 1]$ .

4.5.23. Find an orthogonal basis for the space of solutions to the differential equation  $y''' - y'' + y' - y = 0$  for the  $L^2$  inner product on  $[-\pi, \pi]$ .

- ♡ 4.5.24. In this exercise, we investigate the effect of more general changes of variables on orthogonal polynomials. (a) Prove that  $t = 2s^2 - 1$  defines a one-to-one map from the interval  $0 \leq s \leq 1$  to the interval  $-1 \leq t \leq 1$ . (b) Let  $p_k(t)$  denote the monic Legendre polynomials, which are orthogonal on  $-1 \leq t \leq 1$ . Show that  $q_k(s) = p_k(2s^2 - 1)$  defines a polynomial. Write out the cases  $k = 0, 1, 2, 3$  explicitly. (c) Are the polynomials  $q_k(s)$  orthogonal under the  $L^2$  inner product on  $[0, 1]$ ? If not, do they retain any sort of orthogonality property? *Hint:* What happens to the  $L^2$  inner product on  $[-1, 1]$  under the change of variables?
- 4.5.25. (a) Show that the change of variables  $s = e^{-t}$  maps the Laguerre inner product (4.66) to the standard  $L^2$  inner product on  $[0, 1]$ . However, explain why this does *not* allow you to change Legendre polynomials into Laguerre polynomials. (b) Describe the functions resulting from applying the change of variables to the modified Legendre polynomials (4.74) and their orthogonality properties. (c) Describe the functions that result from applying the inverse change of variables to the Laguerre polynomials (4.68) and their orthogonality properties.
- 4.5.26. Explain how to adapt the numerically stable Gram–Schmidt method in (4.28) to construct a system of orthogonal polynomials. Test your algorithm on one of the preceding exercises.



# Chapter 5

# Minimization and Least Squares

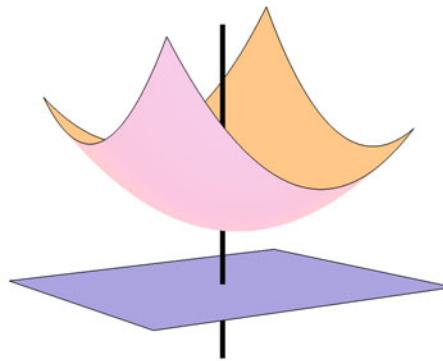
Because Nature seems to strive for efficiency, many systems arising in physical applications are founded on a minimization principle. In a mechanical system, the stable equilibrium configurations minimize the potential energy. In an electrical circuit, the current adjusts itself to minimize the power. In optics and relativity, light rays follow the paths of minimal distance — the geodesics on the curved space-time manifold. Solutions to most of the boundary value problems arising in applications to continuum mechanics are also characterized by a minimization principle, which is then employed to design finite element numerical approximations to their solutions, [61, 81]. Optimization — finding minima or maxima — is ubiquitous throughout mathematical modeling, physics, engineering, economics, and data science, including the calculus of variations, differential geometry, control theory, design and manufacturing, linear programming, machine learning, and beyond.

This chapter introduces and solves the most basic mathematical minimization problem: a quadratic polynomial function depending on several variables. (Minimization of more complicated functions is of comparable significance, but relies on the nonlinear methods of multivariable calculus, and thus lies outside our scope.) Assuming that the quadratic coefficient matrix is positive definite, the minimizer can be found by solving an associated linear algebraic system. Orthogonality also plays an important role in minimization problems. Indeed, the orthogonal projection of a point onto a subspace turns out to be the closest point or least squares minimizer. Moreover, when written in terms of an orthogonal or orthonormal basis for the subspace, the orthogonal projection has an elegant explicit formula that also offers numerical advantages over the direct approach to least squares minimization.

The most common way of fitting a function to prescribed data points is to minimize the least squares error, which serves to quantify the overall deviation between the data and the sampled function values. Our presentation includes an introduction to the interpolation of data points by functions, with a particular emphasis on polynomials and splines. The final Section 5.6 is devoted to the basics of discrete Fourier analysis — the interpolation of data by trigonometric functions — culminating in the remarkable Fast Fourier Transform, a key algorithm in modern signal processing and numerical analysis. Additional applications of these tools in equilibrium mechanics and electrical circuits will form the focus of Chapter 6.

## 5.1 Minimization Problems

Let us begin by introducing three important minimization problems — the first arising in physics, the second in analysis, and the third in geometry.



**Figure 5.1.** Minimizing a Quadratic Function.

## Equilibrium Mechanics

A fundamental principle of mechanics is that systems in equilibrium minimize potential energy. For example, a ball in a bowl will roll downhill unless it is sitting at the bottom, where its potential energy due to gravity is at a (local) minimum. In the simplest class of examples, the energy is a quadratic function, e.g.,

$$f(x, y) = 3x^2 - 2xy + 4y^2 + x - 2y + 1, \quad (5.1)$$

and one seeks the point  $x = x^*$ ,  $y = y^*$ , (if one exists) at which  $f(x^*, y^*)$  achieves its overall minimal value.

Similarly, a pendulum will swing back and forth unless it rests at the bottom of its arc, where potential energy is minimized. Actually, the pendulum has a second equilibrium position at the top of the arc, as in [Figure 5.2](#), but this is rarely observed, since it is an *unstable* equilibrium, meaning that any tiny movement will knock it off balance. Therefore, a better way of stating the principle is that *stable equilibria* are where the mechanical system (locally) minimizes potential energy. For a ball rolling on a curved surface, the local minima — the bottoms of valleys — are the stable equilibria, while the local maxima — the tops of hills — are unstable. Minimization principles serve to characterize the equilibrium configurations of a wide range of physical systems, including masses and springs, structures, electrical circuits, and even continuum models of solid mechanics and elasticity, fluid mechanics, relativity, electromagnetism, thermodynamics, and so on.

## Solution of Equations

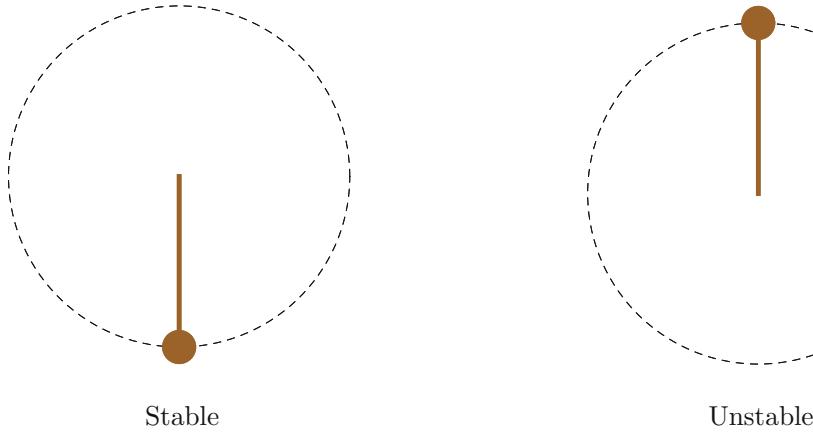
Suppose we wish to solve a system of equations

$$f_1(\mathbf{x}) = 0, \quad f_2(\mathbf{x}) = 0, \quad \dots \quad f_m(\mathbf{x}) = 0, \quad (5.2)$$

where  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ . This system can be converted into a minimization problem in the following seemingly silly manner. Define

$$p(\mathbf{x}) = [f_1(\mathbf{x})]^2 + \dots + [f_m(\mathbf{x})]^2 = \|\mathbf{f}(\mathbf{x})\|^2, \quad (5.3)$$

where  $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^T$  and  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^m$ . Clearly,  $p(\mathbf{x}) \geq 0$  for all  $\mathbf{x}$ . Moreover,  $p(\mathbf{x}^*) = 0$  if and only if each summand is zero, and hence



**Figure 5.2.** Equilibria of a Pendulum.

$\mathbf{x} = \mathbf{x}^*$  is a solution to (5.2). Therefore, the minimum value of  $p(\mathbf{x})$  is zero, and the minimum is achieved if and only if we are at a solution to the original system of equations.

For us, the most important case is that of a linear system

$$A\mathbf{x} = \mathbf{b} \quad (5.4)$$

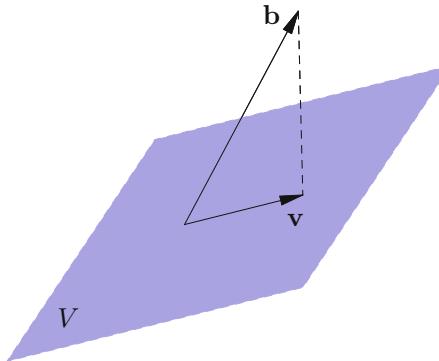
consisting of  $m$  equations in  $n$  unknowns. In this case, the solutions may be obtained by minimizing the function

$$p(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|^2, \quad (5.5)$$

where  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^m$ . Clearly  $p(\mathbf{x})$  has a minimum value of 0, which is achieved if and only if  $\mathbf{x}$  is a solution to the linear system (5.4). Of course, it is not clear that we have gained much, since we already know how to solve  $A\mathbf{x} = \mathbf{b}$  by Gaussian Elimination. However, this artifice turns out to have profound consequences.

Suppose that the linear system (5.4) does *not* have a solution, i.e.,  $\mathbf{b}$  does not lie in the image of the matrix  $A$ . This situation is very typical when there are more equations than unknowns. Such problems arise in data fitting, when the measured data points are all supposed to lie on a straight line, say, but rarely do so exactly, due to experimental error. Although we know there is no exact solution to the system, we might still try to find an approximate solution — a vector  $\mathbf{x}^*$  that comes as close to solving the system as possible. One way to measure closeness is by looking at the magnitude of the *error* as measured by the *residual vector*  $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ , i.e., the difference between the right- and left-hand sides of the system. The smaller its norm  $\|\mathbf{r}\| = \|A\mathbf{x} - \mathbf{b}\|$ , the better the attempted solution. For the Euclidean norm, the vector  $\mathbf{x}^*$  that minimizes the squared residual norm function (5.5) is known as the *least squares solution* to the linear system, because  $\|\mathbf{r}\|^2 = r_1^2 + \cdots + r_n^2$  is the sum of the squares of the individual error components. As before, if the linear system (5.4) happens to have an actual solution, with  $A\mathbf{x}^* = \mathbf{b}$ , then  $\mathbf{x}^*$  qualifies as the least squares solution too, since in this case,  $\|A\mathbf{x}^* - \mathbf{b}\| = 0$  achieves its absolute minimum. So least squares solutions include traditional solutions as special cases.

Unlike an exact solution, the least squares minimizer depends on the choice of inner product governing the norm; thus a suitable weighted norm can be introduced to emphasize or de-emphasize the various errors. While not the only possible approach, least squares is certainly the easiest to analyze and solve, and, hence, is often the method of choice for



**Figure 5.3.** The Closest Point.

fitting functions to experimental data and performing statistical analysis. It is essential that the norm arise from an inner product; minimizing the error based on other kinds of norms is a much more difficult, nonlinear problem, although one that has recently become of immense practical interest in the newly emergent field of compressed sensing, [28].

### The Closest Point

The following minimization problem arises in elementary geometry, although its practical implications cut a much wider swath. Given a point  $\mathbf{b} \in \mathbb{R}^m$  and a subset  $V \subset \mathbb{R}^m$ , find the point  $\mathbf{v}^* \in V$  that is closest to  $\mathbf{b}$ . In other words, we seek to minimize the Euclidean distance  $d(\mathbf{v}, \mathbf{b}) = \|\mathbf{v} - \mathbf{b}\|$  over all possible  $\mathbf{v} \in V$ .

The simplest situation occurs when  $V$  is a subspace of  $\mathbb{R}^m$ . In this case, the closest point problem can, in fact, be reformulated as a least squares minimization problem. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis for  $V$ . The general element  $\mathbf{v} \in V$  is a linear combination of the basis vectors. Applying our handy matrix multiplication formula (2.13), we can write the subspace elements in the form

$$\mathbf{v} = x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n = A \mathbf{x},$$

where  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  is the  $m \times n$  matrix formed by the (column) basis vectors and  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  are the coordinates of  $\mathbf{v}$  relative to the chosen basis. In this manner, we can identify  $V$  with the image of  $A$ , i.e., the subspace spanned by its columns. Consequently, the closest point in  $V$  to  $\mathbf{b}$  is found by minimizing  $\|\mathbf{v} - \mathbf{b}\|^2 = \|A \mathbf{x} - \mathbf{b}\|^2$  over all possible  $\mathbf{x} \in \mathbb{R}^n$ . But this is exactly the same as the least squares function (5.5)! Thus, if  $\mathbf{x}^*$  is the least squares solution to the system  $A \mathbf{x} = \mathbf{b}$ , then  $\mathbf{v}^* = A \mathbf{x}^*$  is the closest point to  $\mathbf{b}$  belonging to  $V = \text{img } A$ . In this way, we have established a profound and fertile connection between least squares solutions to linear systems and the geometrical problem of minimizing distances to subspaces. And, as we shall see, the closest point  $\mathbf{v} \in V$  turns out to be the orthogonal projection of  $\mathbf{b}$  onto the subspace.

All three of the preceding minimization problems are solved by the same underlying mathematical construction, which will now be described in detail.

**Remark.** In this book, we will concentrate on minimization problems. Maximizing a function  $p(\mathbf{x})$  is the same as minimizing its negative  $-p(\mathbf{x})$ , and so can be easily handled by the same techniques.

---

## Exercises

*Note:* Unless otherwise indicated, “distance” refers to the Euclidean norm.

- 5.1.1. Find the least squares solution to the pair of equations  $3x = 1$ ,  $2x = -1$ .
- 5.1.2. Find the minimizer of the function  $f(x, y) = (3x - 2y + 1)^2 + (2x + y + 2)^2$ .
- 5.1.3. Find the closest point or points to  $\mathbf{b} = (-1, 2)^T$  that lie on (a) the  $x$ -axis, (b) the  $y$ -axis, (c) the line  $y = x$ , (d) the line  $x + y = 0$ , (e) the line  $2x + y = 0$ .
- 5.1.4. Solve Exercise 5.1.3 when distance is measured in (i) the  $\infty$  norm, (ii) the 1 norm.
- 5.1.5. Given  $\mathbf{b} \in \mathbb{R}^2$ , is the closest point on a line  $L$  unique when distance is measured in (a) the Euclidean norm? (b) the 1 norm? (c) the  $\infty$  norm?
- ♡ 5.1.6. Let  $L \subset \mathbb{R}^2$  be a line through the origin, and let  $\mathbf{b} \in \mathbb{R}^2$  be any point.
- Find a geometrical construction of the closest point  $\mathbf{v} \in L$  to  $\mathbf{b}$  when distance is measured in the standard Euclidean norm.
  - Use your construction to prove that there is one and only one closest point.
- (c) Show that if  $\mathbf{0} \neq \mathbf{a} \in L$ , then the distance equals  $\frac{\sqrt{\|\mathbf{a}\|^2 \|\mathbf{b}\|^2 - (\mathbf{a} \cdot \mathbf{b})^2}}{\|\mathbf{a}\|} = \frac{|\mathbf{a} \times \mathbf{b}|}{\|\mathbf{a}\|}$ , using the two-dimensional cross product (3.22).
- 5.1.7. Suppose  $\mathbf{a}$  and  $\mathbf{b}$  are unit vectors in  $\mathbb{R}^2$ . Show that the distance from  $\mathbf{a}$  to the line through  $\mathbf{b}$  is the same as the distance from  $\mathbf{b}$  to the line through  $\mathbf{a}$ . Use a picture to explain why this holds. How is the distance related to the angle between the two vectors?
- 5.1.8. (a) Prove that the distance from the point  $(x_0, y_0)^T$  to the line  $ax + by = 0$  is  $\frac{|ax_0 + by_0|}{\sqrt{a^2 + b^2}}$ . (b) What is the minimum distance to the line  $ax + by + c = 0$ ?
- ♡ 5.1.9. (a) Generalize Exercise 5.1.8 to find the distance between a point  $(x_0, y_0, z_0)^T$  and the plane  $ax + by + cz + d = 0$  in  $\mathbb{R}^3$ . (b) Use your formula to compute the distance between  $(1, 1, 1)^T$  and the plane  $3x - 2y + z = 1$ .
- 5.1.10. (a) Explain in detail why the minimizer of  $\|\mathbf{v} - \mathbf{b}\|$  coincides with the minimizer of  $\|\mathbf{v} - \mathbf{b}\|^2$ . (b) Find all scalar functions  $F(x)$  for which the minimizer of  $F(\|\mathbf{v} - \mathbf{b}\|)$  is the same as the minimizer of  $\|\mathbf{v} - \mathbf{b}\|$ .
- 5.1.11. (a) Explain why the problem of maximizing the distance from a point to a subspace does not have a solution. (b) Can you formulate a situation in which maximizing distance to a point leads to a problem with a solution?

---

## 5.2 Minimization of Quadratic Functions

The simplest algebraic equations are linear systems. As such, one must thoroughly understand them before venturing into the far more complicated nonlinear realm. For minimization problems, the starting point is the quadratic function. (Linear functions do not have minima — think of the function  $f(x) = \alpha x + \beta$ , whose graph is a straight line<sup>†</sup>.) In this section, we shall see how the problem of minimizing a general quadratic function of  $n$

---

<sup>†</sup> Technically, this function is linear only when  $\beta = 0$ ; otherwise it is known as an “affine function”. See Chapter 7 for details.

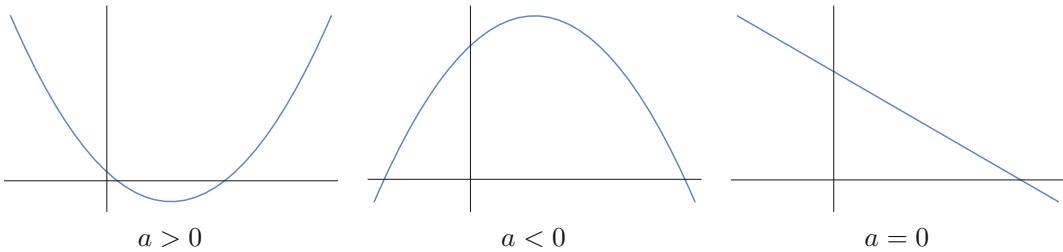


Figure 5.4. Parabolas.

variables can be solved by linear algebra techniques.

Let us begin by reviewing the very simplest example — minimizing a scalar quadratic polynomial

$$p(x) = ax^2 + 2bx + c \quad (5.6)$$

over all possible values of  $x \in \mathbb{R}$ . If  $a > 0$ , then the graph of  $p$  is a parabola opening upwards, and so there exists a unique minimum value. If  $a < 0$ , the parabola points downwards, and there is no minimum (although there is a maximum). If  $a = 0$ , the graph is a straight line, and there is neither minimum nor maximum over all  $x \in \mathbb{R}$  — except in the trivial case  $b = 0$  also, and the function  $p(x) = c$  is constant, with every  $x$  qualifying as a minimum (and a maximum). The three nontrivial possibilities are illustrated in Figure 5.4.

In the case  $a > 0$ , the minimum can be found by calculus. The *critical points* of a function, which are candidates for minima (and maxima), are found by setting its derivative to zero. In this case, differentiating, and solving

$$p'(x) = 2ax + 2b = 0,$$

we conclude that the only possible minimum value occurs at

$$x^* = -\frac{b}{a}, \quad \text{where} \quad p(x^*) = c - \frac{b^2}{a}. \quad (5.7)$$

Of course, one must check that this critical point is indeed a minimum, and not a maximum or inflection point. The second derivative test will show that  $p''(x^*) = 2a > 0$ , and so  $x^*$  is at least a local minimum.

A more instructive approach to this problem — and one that requires only elementary algebra — is to “complete the square”. As in (3.66), we rewrite

$$p(x) = a \left( x + \frac{b}{a} \right)^2 + \frac{ac - b^2}{a}. \quad (5.8)$$

If  $a > 0$ , then the first term is always  $\geq 0$ , and, moreover, attains its minimum value 0 only at  $x^* = -b/a$ . The second term is constant, and so is unaffected by the value of  $x$ . Thus, the *global minimum* of  $p(x)$  is at  $x^* = -b/a$ . Moreover, its minimal value equals the constant term,  $p(x^*) = c - b^2/a$ , thereby reconfirming and strengthening the calculus result in (5.7).

Now that we have the one-variable case firmly in hand, let us turn our attention to the more substantial problem of minimizing quadratic functions of several variables. Thus, we seek to minimize a (real) *quadratic polynomial*

$$p(\mathbf{x}) = p(x_1, \dots, x_n) = \sum_{i,j=1}^n k_{ij} x_i x_j - 2 \sum_{i=1}^n f_i x_i + c, \quad (5.9)$$

depending on  $n$  variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ . The coefficients  $k_{ij}, f_i$  and  $c$  are all assumed to be real. Moreover, we can assume, without loss of generality, that the coefficients of the quadratic terms are symmetric:  $k_{ij} = k_{ji}$ . (See Exercise 3.4.15 for a justification.) Note that  $p(\mathbf{x})$  is more general than a quadratic form (3.52) in that it also contains linear and constant terms. We seek a global minimum, and so the variables  $\mathbf{x}$  are allowed to vary over all of  $\mathbb{R}^n$ . (Minimizing a quadratic function over a proper subset  $\mathbf{x} \in S \subsetneq \mathbb{R}^n$  is a more challenging problem, and will not be discussed here.)

Let us begin by rewriting the quadratic function (5.9) in a more compact matrix notation:

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T \mathbf{f} + c, \quad \mathbf{x} \in \mathbb{R}^n, \quad (5.10)$$

in which  $K = (k_{ij})$  is a symmetric  $n \times n$  matrix,  $\mathbf{f} \in \mathbb{R}^n$  is a constant vector, and  $c$  is a constant scalar.

**Example 5.1.** Consider the quadratic function

$$p(x_1, x_2) = 4x_1^2 - 2x_1x_2 + 3x_2^2 + 3x_1 - 2x_2 + 1$$

depending on two real variables  $x_1, x_2$ . It can be written in the matrix form (5.10) as

$$p(x_1, x_2) = (x_1 \ x_2) \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2(x_1 \ x_2) \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix} + 1, \quad (5.11)$$

whereby

$$K = \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix}, \quad c = 1. \quad (5.12)$$

Pay attention to the symmetry of  $K = K^T$ , whereby its corresponding off-diagonal entries, here both  $-1$ , are each one-half the coefficient of the corresponding quadratic monomial, in this case  $-2x_1x_2$ . Also note the overall factor of  $-2$  in front of the linear terms, which is included for later convenience.

We first note that in the simple scalar case (5.6), we needed to impose the condition that the quadratic coefficient  $a$  be *positive* in order to obtain a (unique) minimum. The corresponding condition for the multivariable case is that the quadratic coefficient matrix  $K$  be *positive definite*. This key assumption enables us to establish a general minimization criterion.

**Theorem 5.2.** If  $K$  is a positive definite (and hence symmetric) matrix, then the quadratic function (5.10) has a unique minimizer, which is the solution to the linear system

$$K \mathbf{x} = \mathbf{f}, \quad \text{namely} \quad \mathbf{x}^* = K^{-1} \mathbf{f}. \quad (5.13)$$

The minimum value of  $p(\mathbf{x})$  is equal to any of the following expressions:

$$p(\mathbf{x}^*) = p(K^{-1} \mathbf{f}) = c - \mathbf{f}^T K^{-1} \mathbf{f} = c - \mathbf{f}^T \mathbf{x}^* = c - (\mathbf{x}^*)^T K \mathbf{x}^*. \quad (5.14)$$

*Proof:* First recall that, by Proposition 3.31, positive definiteness implies that  $K$  is a nonsingular matrix, and hence the linear system (5.13) has a unique solution  $\mathbf{x}^* = K^{-1} \mathbf{f}$ . Then, for all  $\mathbf{x} \in \mathbb{R}^n$ , since  $\mathbf{f} = K \mathbf{x}^*$ , it follows that

$$\begin{aligned} p(\mathbf{x}) &= \mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T \mathbf{f} + c = \mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T K \mathbf{x}^* + c \\ &= (\mathbf{x} - \mathbf{x}^*)^T K (\mathbf{x} - \mathbf{x}^*) + [c - (\mathbf{x}^*)^T K \mathbf{x}^*], \end{aligned} \quad (5.15)$$

where we used the symmetry of  $K = K^T$  to identify the scalar terms

$$\mathbf{x}^T K \mathbf{x}^* = (\mathbf{x}^T K \mathbf{x}^*)^T = (\mathbf{x}^*)^T K^T \mathbf{x} = (\mathbf{x}^*)^T K \mathbf{x}.$$

The first term in the final expression in (5.15) has the form  $\mathbf{y}^T K \mathbf{y}$ , where  $\mathbf{y} = \mathbf{x} - \mathbf{x}^*$ . Since we assumed that  $K$  is positive definite, we know that  $\mathbf{y}^T K \mathbf{y} > 0$  for all  $\mathbf{y} \neq \mathbf{0}$ . Thus, the first term achieves its minimum value, namely 0, if and only if  $\mathbf{0} = \mathbf{y} = \mathbf{x} - \mathbf{x}^*$ . Since  $\mathbf{x}^*$  is fixed, the second, bracketed, term does not depend on  $\mathbf{x}$ , and hence the minimizer of  $p(\mathbf{x})$  coincides with the minimizer of the first term, namely  $\mathbf{x} = \mathbf{x}^*$ . Moreover, the minimum value of  $p(\mathbf{x})$  is equal to the constant term:  $p(\mathbf{x}^*) = c - (\mathbf{x}^*)^T K \mathbf{x}^*$ . The alternative expressions in (5.14) follow from simple substitutions. *Q.E.D.*

**Example 5.1 (continued).** Let us minimize the quadratic function appearing in (5.11) above. According to Theorem 5.2, to find the minimum we must solve the linear system  $K \mathbf{x} = \mathbf{f}$ , which, in this case, is

$$\begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix}. \quad (5.16)$$

When applying the usual Gaussian Elimination algorithm, only one row operation is required to place the coefficient matrix in upper triangular form:

$$\begin{pmatrix} 4 & -1 & | & -\frac{3}{2} \\ -1 & 3 & | & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 4 & -1 & | & -\frac{3}{2} \\ 0 & \frac{11}{4} & | & \frac{5}{8} \end{pmatrix}.$$

The coefficient matrix is regular, since no row interchanges were required, and its two pivots, namely 4 and  $\frac{11}{4}$ , are both positive. Thus, by Theorem 3.43,  $K > 0$ , and hence  $p(x_1, x_2)$  really does have a minimum, obtained by applying Back Substitution to the reduced system:

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = \begin{pmatrix} -\frac{7}{22} \\ \frac{5}{22} \end{pmatrix} \approx \begin{pmatrix} -.31818 \\ .22727 \end{pmatrix}. \quad (5.17)$$

The quickest way to compute the minimal value is to use the second formula in (5.14):

$$p(\mathbf{x}^*) = p\left(-\frac{7}{22}, \frac{5}{22}\right) = 1 - \left(-\frac{3}{2}, 1\right) \begin{pmatrix} -\frac{7}{22} \\ \frac{5}{22} \end{pmatrix} = \frac{13}{44} \approx .29546$$

It is instructive to compare the algebraic solution method with the minimization procedure you learned in multi-variable calculus, cf. [2, 78]. The *critical points* of  $p(x_1, x_2)$  are found by setting both partial derivatives equal to zero:

$$\frac{\partial p}{\partial x_1} = 8x_1 - 2x_2 + 3 = 0, \quad \frac{\partial p}{\partial x_2} = -2x_1 + 6x_2 - 2 = 0.$$

If we divide by an overall factor of 2, these are precisely the *same* linear equations we already constructed in (5.16). Thus, not surprisingly, the calculus approach leads to the same minimizer (5.17). To check whether  $\mathbf{x}^*$  is a (local) minimum, we need to apply the second derivative test. In the case of a function of several variables, this requires analyzing the *Hessian matrix*, which is the symmetric matrix of second order partial derivatives

$$H = \begin{pmatrix} \frac{\partial^2 p}{\partial x_1^2} & \frac{\partial^2 p}{\partial x_1 \partial x_2} \\ \frac{\partial^2 p}{\partial x_1 \partial x_2} & \frac{\partial^2 p}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} 8 & -2 \\ -2 & 6 \end{pmatrix} = 2K,$$

which is exactly twice the quadratic coefficient matrix (5.12). If the Hessian matrix is positive definite — which we already know in this case — then the critical point is indeed a (local) minimum.

Thus, the calculus and algebraic approaches to this minimization problem lead, as they must, to identical results. However, the algebraic method is *more* powerful, because it immediately produces the *unique, global* minimum, whereas, barring additional work, calculus can guarantee only that the critical point is a local minimum. Moreover, the proof of the calculus local minimization criterion — that the Hessian matrix be positive definite at the critical point — relies, in fact, on the algebraic solution to the quadratic minimization problem! In summary: minimization of quadratic functions is a problem in linear algebra, while minimizing more complicated functions requires the full force of multivariable calculus.

The most efficient method for producing a minimum of a quadratic function  $p(\mathbf{x})$  on  $\mathbb{R}^n$ , then, is to first write out the symmetric coefficient matrix  $K$  and the vector  $\mathbf{f}$  as in (5.10). Solving the system  $K\mathbf{x} = \mathbf{f}$  will produce the minimizer  $\mathbf{x}^*$  *provided*  $K > 0$  — which should be checked during the course of the procedure using the criteria of Theorem 3.43, that is, making sure that no row interchanges are used and all the pivots are positive.

**Example 5.3.** Let us minimize the quadratic function

$$p(x, y, z) = x^2 + 2xy + xz + 2y^2 + yz + 2z^2 + 6y - 7z + 5.$$

This has the matrix form (5.10) with

$$K = \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 0 \\ -3 \\ \frac{7}{2} \end{pmatrix}, \quad c = 5,$$

and the minimum is found by solving the linear system,  $K\mathbf{x} = \mathbf{f}$ . Gaussian Elimination produces the  $LDL^T$  factorization

$$K = \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{7}{4} \end{pmatrix} \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The pivots, i.e., the diagonal entries of  $D$ , are all positive, and hence  $K$  is positive definite. Theorem 5.2 then guarantees that  $p(x, y, z)$  has a unique minimizer, which is found by solving the linear system  $K\mathbf{x} = \mathbf{f}$ . The solution is then quickly obtained by forward and back substitution:

$$x^* = 2, \quad y^* = -3, \quad z^* = 2, \quad \text{with} \quad p(x^*, y^*, z^*) = p(2, -3, 2) = -11,$$

and we conclude that  $p(x, y, z) > p(2, -3, 2) = -11$  for all  $(x, y, z) \neq (2, -3, 2)$ .

Theorem 5.2 solves the general quadratic minimization problem when the quadratic coefficient matrix is positive definite. If  $K$  is not positive definite, then the quadratic function (5.10) does not have a minimum, apart from one exceptional situation.

**Theorem 5.4.** If the matrix  $K$  is positive definite, then the quadratic function (5.10) has a unique global minimizer  $\mathbf{x}^*$  satisfying  $K\mathbf{x}^* = \mathbf{f}$ . If  $K$  is only positive semi-definite, and  $\mathbf{f} \in \text{img } K$ , then every solution to the linear system  $K\mathbf{x}^* = \mathbf{f}$  is a global minimum of  $p(\mathbf{x})$ , but the minimum is not unique, since  $p(\mathbf{x}^* + \mathbf{z}) = p(\mathbf{x}^*)$  whenever  $\mathbf{z} \in \ker K$ . In all other cases,  $p(\mathbf{x})$  has no global minimum.

*Proof:* The first part is merely a restatement of Theorem 5.2. The second part is proved by a similar computation, and is left to the reader. If  $K$  is not positive semi-definite, then one can find a vector  $\mathbf{y}$  such that  $a = \mathbf{y}^T K \mathbf{y} < 0$ . If we set  $\mathbf{x} = t\mathbf{y}$ , then  $p(\mathbf{x}) = p(t\mathbf{y}) = at^2 + 2bt + c$ , with  $b = \mathbf{y}^T \mathbf{f}$ . Since  $a < 0$ , by choosing  $|t| \gg 0$  sufficiently large, we can arrange that  $p(t\mathbf{y}) \ll 0$  is an arbitrarily large negative quantity, and so  $p$  has no (finite) minimum value. The one remaining case — when  $K$  is positive semi-definite, but  $\mathbf{f} \notin \text{img } K$  — is the subject of Exercise 5.2.14. *Q.E.D.*

## Exercises

5.2.1. Find the minimum value of the function  $f(x, y, z) = x^2 + 2xy + 3y^2 + 2yz + z^2 - 2x + 3z + 2$ . How do you know that your answer is really the global minimum?

5.2.2. For the potential energy function in (5.1), where is the equilibrium position of the ball?

5.2.3. For each of the following quadratic functions, determine whether there is a minimum. If so, find the minimizer and the minimum value for the function.

- (a)  $x^2 - 2xy + 4y^2 + x - 1$ , (b)  $3x^2 + 3xy + 3y^2 - 2x - 2y + 4$ , (c)  $x^2 + 5xy + 3y^2 + 2x - y$ ,  
 (d)  $x^2 + y^2 + yz + z^2 + x + y - z$ , (e)  $x^2 + xy - y^2 - yz + z^2 - 3$ ,  
 (f)  $x^2 + 5xz + y^2 - 2yz + z^2 + 2x - z - 3$ , (g)  $x^2 + xy + y^2 + yz + z^2 + zw + w^2 - 2x - w$ .

5.2.4. (a) For which numbers  $b$  (allowing both positive and negative numbers) is the matrix

$A = \begin{pmatrix} 1 & b \\ b & 4 \end{pmatrix}$  positive definite? (b) Find the factorization  $A = LDL^T$  when  $b$  is in the range for positive definiteness. (c) Find the minimum value (depending on  $b$ ; it might be finite or it might be  $-\infty$ ) of the function  $p(x, y) = x^2 + 2bxy + 4y^2 - 2y$ .

5.2.5. For each matrix  $K$ , vector  $\mathbf{f}$ , and scalar  $c$ , write out the quadratic function  $p(\mathbf{x})$  given by (5.10). Then either find the minimizer  $\mathbf{x}^*$  and minimum value  $p(\mathbf{x}^*)$ , or explain why there

- is none. (a)  $K = \begin{pmatrix} 4 & -12 \\ -12 & 45 \end{pmatrix}$ ,  $\mathbf{f} = \begin{pmatrix} -\frac{1}{2} \\ 2 \end{pmatrix}$ ,  $c = 3$ ; (b)  $K = \begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix}$ ,  $\mathbf{f} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$ ,  
 $c = 0$ ; (c)  $K = \begin{pmatrix} 3 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 3 \end{pmatrix}$ ,  $\mathbf{f} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$ ,  $c = -3$ ; (d)  $K = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -1 \\ 1 & -1 & 1 \end{pmatrix}$ ,  
 $\mathbf{f} = \begin{pmatrix} -3 \\ -1 \\ 2 \end{pmatrix}$ ,  $c = 1$ ; (e)  $K = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}$ ,  $\mathbf{f} = \begin{pmatrix} -1 \\ 2 \\ -3 \\ 4 \end{pmatrix}$ ,  $c = 0$ .

5.2.6. Find the minimum value of the quadratic function

$$p(x_1, \dots, x_n) = 4 \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^{n-1} x_i x_{i+1} + \sum_{i=1}^n x_i \quad \text{for } n = 2, 3, 4.$$

5.2.7. Find the maximum value of the quadratic functions

- (a)  $-x^2 + 3xy - 5y^2 - x + 1$ , (b)  $-2x^2 + 6xy - 3y^2 + 4x - 3y$ .

5.2.8. Suppose  $K_1$  and  $K_2$  are positive definite  $n \times n$  matrices. Suppose that, for  $i = 1, 2$ , the minimizer of  $p_i(\mathbf{x}) = \mathbf{x}^T K_i \mathbf{x} - 2\mathbf{x}^T \mathbf{f}_i + c_i$ , is  $\mathbf{x}_i^*$ . Is the minimizer of  $p(\mathbf{x}) = p_1(\mathbf{x}) + p_2(\mathbf{x})$  given by  $\mathbf{x}^* = \mathbf{x}_1^* + \mathbf{x}_2^*$ ? Prove or give a counterexample.

◇ 5.2.9. Let  $K > 0$ . Prove that a quadratic function  $p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f}$  without constant term has non-positive minimum value:  $p(\mathbf{x}^*) \leq 0$ . When is the minimum value zero?

5.2.10. Let  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$  be a quadratic form. Prove that the minimum value of  $q(\mathbf{x})$  is either 0 or  $-\infty$ .

5.2.11. Under what conditions does the affine function  $p(\mathbf{x}) = \mathbf{x}^T \mathbf{f} + c$  have a minimum?

◇ 5.2.12. Under what conditions does a quadratic function  $p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c$  have a finite global maximum? Explain how to find the maximizer and maximum value.

5.2.13. *True or false:* The minimal-norm solution to  $A\mathbf{x} = \mathbf{b}$  is obtained by setting all the free variables to zero.

◇ 5.2.14. Prove that if  $K$  is a positive semi-definite matrix, and  $\mathbf{f} \notin \text{img } K$ , then the quadratic function  $p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c$  has no minimum value.

*Hint:* Try looking at vectors  $\mathbf{x} \in \ker K$ .

5.2.15. Why can't you minimize a complex-valued quadratic function?

### 5.3 The Closest Point

We are now ready to solve the geometric problem of finding the element in a prescribed subspace that lies closest to a given point. For simplicity, we work mostly with subspaces of  $\mathbb{R}^m$ , equipped with the Euclidean norm and inner product, but the method extends straightforwardly to arbitrary finite-dimensional subspaces of any inner product space. However, it does *not* apply to more general norms not associated with inner products, such as the 1 norm, the  $\infty$  norm and, in fact, the  $p$  norms whenever  $p \neq 2$ . In such cases, finding the closest point problem is a *nonlinear* minimization problem whose solution requires more sophisticated analytical techniques; see, for example, [28, 66, 79].

**Problem.** Let  $\mathbb{R}^m$  be equipped with an inner product  $\langle \mathbf{v}, \mathbf{w} \rangle$  and associated norm  $\|\mathbf{v}\|$ , and let  $W \subset \mathbb{R}^m$  be a subspace. Given  $\mathbf{b} \in \mathbb{R}^m$ , the goal is to find the point  $\mathbf{w}^* \in W$  that minimizes  $\|\mathbf{w} - \mathbf{b}\|$  over all possible  $\mathbf{w} \in W$ . The minimal distance  $d^* = \|\mathbf{w}^* - \mathbf{b}\|$  to the closest point is designated as the *distance* from the point  $\mathbf{b}$  to the subspace  $W$ .

Of course, if  $\mathbf{b} \in W$  lies in the subspace, then the answer is easy: the closest point in  $W$  is  $\mathbf{w}^* = \mathbf{b}$  itself, and the distance from  $\mathbf{b}$  to the subspace is zero. Thus, the problem becomes interesting only when  $\mathbf{b} \notin W$ .

In solving the closest point problem, the goal is to minimize the squared distance

$$\|\mathbf{w} - \mathbf{b}\|^2 = \langle \mathbf{w} - \mathbf{b}, \mathbf{w} - \mathbf{b} \rangle = \|\mathbf{w}\|^2 - 2\langle \mathbf{w}, \mathbf{b} \rangle + \|\mathbf{b}\|^2 \quad (5.18)$$

over all possible  $\mathbf{w}$  belonging to the subspace  $W \subset \mathbb{R}^m$ . Let us assume that we know a basis  $\mathbf{w}_1, \dots, \mathbf{w}_n$  of  $W$ , with  $n = \dim W$ . Then the most general vector in  $W$  is a linear combination

$$\mathbf{w} = x_1 \mathbf{w}_1 + \dots + x_n \mathbf{w}_n \quad (5.19)$$

of the basis vectors. We substitute the formula (5.19) for  $\mathbf{w}$  into the squared distance function (5.18). As we shall see, the resulting expression is a quadratic function of the coefficients  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ , and so the minimum is provided by Theorem 5.2.

First, the quadratic terms come from expanding

$$\|\mathbf{w}\|^2 = \langle x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n, x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n \rangle = \sum_{i,j=1}^n x_i x_j \langle \mathbf{w}_i, \mathbf{w}_j \rangle. \quad (5.20)$$

Therefore,

$$\|\mathbf{w}\|^2 = \sum_{i,j=1}^n k_{ij} x_i x_j = \mathbf{x}^T K \mathbf{x},$$

where  $K$  is the symmetric  $n \times n$  *Gram matrix* whose  $(i,j)$  entry is the inner product

$$k_{ij} = \langle \mathbf{w}_i, \mathbf{w}_j \rangle \quad (5.21)$$

between the basis vectors of our subspace; see Definition 3.33. Similarly,

$$\langle \mathbf{w}, \mathbf{b} \rangle = \langle x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n, \mathbf{b} \rangle = \sum_{i=1}^n x_i \langle \mathbf{w}_i, \mathbf{b} \rangle,$$

and so

$$\langle \mathbf{w}, \mathbf{b} \rangle = \sum_{i=1}^n x_i f_i = \mathbf{x}^T \mathbf{f},$$

where  $\mathbf{f} \in \mathbb{R}^n$  is the vector whose  $i^{\text{th}}$  entry is the inner product

$$f_i = \langle \mathbf{w}_i, \mathbf{b} \rangle \quad (5.22)$$

between the point and the subspace's basis elements. Substituting back, we conclude that the squared distance function (5.18) reduces to the quadratic function

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T \mathbf{f} + c = \sum_{i,j=1}^n k_{ij} x_i x_j - 2 \sum_{i=1}^n f_i x_i + c, \quad (5.23)$$

in which  $K$  and  $\mathbf{f}$  are given in (5.21–22), while  $c = \|\mathbf{b}\|^2$ .

Since we assumed that the basis vectors  $\mathbf{w}_1, \dots, \mathbf{w}_n$  are linearly independent, Proposition 3.36 assures us that their associated Gram matrix is positive definite. Therefore, we may directly apply our basic Minimization Theorem 5.2 to solve the closest point problem.

**Theorem 5.5.** Let  $\mathbf{w}_1, \dots, \mathbf{w}_n$  form a basis for the subspace  $W \subset \mathbb{R}^m$ . Given  $\mathbf{b} \in \mathbb{R}^m$ , the closest point  $\mathbf{w}^* = x_1^* \mathbf{w}_1 + \cdots + x_n^* \mathbf{w}_n \in W$  is unique and prescribed by the solution  $\mathbf{x}^* = K^{-1} \mathbf{f}$  to the linear system

$$K \mathbf{x} = \mathbf{f}, \quad (5.24)$$

where the entries of  $K$  and  $\mathbf{f}$  are given in (5.21–22). The (minimum) distance between the point and the subspace is

$$d^* = \|\mathbf{w}^* - \mathbf{b}\| = \sqrt{\|\mathbf{b}\|^2 - \mathbf{f}^T \mathbf{x}^*}. \quad (5.25)$$

When the standard dot product and Euclidean norm on  $\mathbb{R}^m$  are used to measure distance, the entries of the Gram matrix  $K$  and the vector  $\mathbf{f}$  are given by

$$k_{ij} = \mathbf{w}_i \cdot \mathbf{w}_j = \mathbf{w}_i^T \mathbf{w}_j, \quad f_i = \mathbf{w}_i \cdot \mathbf{b} = \mathbf{w}_i^T \mathbf{b}.$$

As in (3.62), each set of equations can be combined into a single matrix equation. If  $A = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n)$  denotes the  $m \times n$  matrix formed by the basis vectors, then

$$K = A^T A, \quad \mathbf{f} = A^T \mathbf{b}, \quad c = \mathbf{b}^T \mathbf{b} = \|\mathbf{b}\|^2. \quad (5.26)$$

A direct derivation of these equations is instructive. Since, by formula (2.13),

$$\mathbf{w} = x_1 \mathbf{w}_1 + \cdots + x_n \mathbf{w}_n = A \mathbf{x},$$

we have

$$\begin{aligned} \|\mathbf{w} - \mathbf{b}\|^2 &= \|A \mathbf{x} - \mathbf{b}\|^2 = (A \mathbf{x} - \mathbf{b})^T (A \mathbf{x} - \mathbf{b}) = (\mathbf{x}^T A^T - \mathbf{b}^T) (A \mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^T A^T A \mathbf{x} - 2 \mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b} = \mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T \mathbf{f} + c, \end{aligned}$$

thereby justifying (5.26). Thus, Theorem 5.5 implies that the closest point  $\mathbf{w}^* = A \mathbf{x}^* \in W$  to  $\mathbf{b}$  in the Euclidean norm is obtained by solving what are known as the *normal equations*

$$(A^T A) \mathbf{x} = A^T \mathbf{b}, \quad (5.27)$$

for

$$\mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{b}, \quad \text{giving} \quad \mathbf{w}^* = A \mathbf{x}^* = A (A^T A)^{-1} A^T \mathbf{b}. \quad (5.28)$$

If, instead of the Euclidean inner product, we adopt a weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w}$  on  $\mathbb{R}^m$  prescribed by a positive definite  $m \times m$  matrix  $C > 0$ , then the same computations produce

$$K = A^T C A, \quad \mathbf{f} = A^T C \mathbf{b}, \quad c = \mathbf{b}^T C \mathbf{b} = \|\mathbf{b}\|^2. \quad (5.29)$$

The resulting formula for the weighted Gram matrix  $K$  was previously derived in (3.64). In this case, the closest point  $\mathbf{w}^* \in W$  in the weighted norm is obtained by solving the *weighted normal equations*

$$A^T C A \mathbf{x} = A^T C \mathbf{b}, \quad (5.30)$$

so that

$$\mathbf{x}^* = (A^T C A)^{-1} A^T C \mathbf{b}, \quad \mathbf{w}^* = A \mathbf{x}^* = A (A^T C A)^{-1} A^T C \mathbf{b}. \quad (5.31)$$

**Example 5.6.** Let  $W \subset \mathbb{R}^3$  be the plane spanned by  $\mathbf{w}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ ,  $\mathbf{w}_2 = \begin{pmatrix} 2 \\ -3 \\ -1 \end{pmatrix}$ . Our goal is to find the point  $\mathbf{w}^* \in W$  closest to  $\mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ , where distance is measured in the usual Euclidean norm. We combine the basis vectors to form the matrix  $A = \begin{pmatrix} 1 & 2 \\ 2 & -3 \\ -1 & -1 \end{pmatrix}$ . According to (5.26), the positive definite Gram matrix and associated vector are

$$K = A^T A = \begin{pmatrix} 6 & -3 \\ -3 & 14 \end{pmatrix}, \quad \mathbf{f} = A^T \mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

(Alternatively, these can be computed directly by taking inner products, as in (5.21–22).) We solve the linear system

$$K \mathbf{x} = \mathbf{f} \quad \text{for} \quad \mathbf{x}^* = K^{-1} \mathbf{f} = \begin{pmatrix} \frac{4}{15} \\ \frac{1}{5} \end{pmatrix}.$$

Theorem 5.5 implies that the closest point is

$$\mathbf{w}^* = A \mathbf{x}^* = x_1^* \mathbf{w}_1 + x_2^* \mathbf{w}_2 = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{15} \\ -\frac{7}{15} \end{pmatrix} \approx \begin{pmatrix} .6667 \\ -.0667 \\ -.4667 \end{pmatrix}.$$

The distance from the point  $\mathbf{b}$  to the plane is  $d^* = \|\mathbf{w}^* - \mathbf{b}\| = \frac{1}{\sqrt{3}} \approx .5774$ .

Suppose, on the other hand, that distance in  $\mathbb{R}^3$  is measured in the weighted norm  $\|\mathbf{v}\| = v_1^2 + \frac{1}{2}v_2^2 + \frac{1}{3}v_3^2$  corresponding to the positive definite diagonal matrix  $C = \text{diag}(1, \frac{1}{2}, \frac{1}{3})$ . In this case, we form the weighted Gram matrix and vector (5.29):

$$K = A^T C A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & -3 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} \frac{10}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{53}{6} \end{pmatrix},$$

$$\mathbf{f} = A^T C \mathbf{b} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

and so

$$\mathbf{x}^* = K^{-1} \mathbf{f} \approx \begin{pmatrix} .3506 \\ .2529 \end{pmatrix}, \quad \mathbf{w}^* = A \mathbf{x}^* \approx \begin{pmatrix} .8563 \\ -.0575 \\ -.6034 \end{pmatrix}.$$

Now the distance between the point and the subspace is measured in the weighted norm:  $d^* = \|\mathbf{w}^* - \mathbf{b}\| \approx .3790$ .

**Remark.** The solution to the closest point problem given in Theorem 5.5 applies, as stated, to the more general case in which  $W \subset V$  is a finite-dimensional subspace of a general inner product space  $V$ . The underlying inner product space  $V$  can even be infinite-dimensional, as, for example, in least squares approximations in function space.

Now, consider what happens if we know an orthonormal basis  $\mathbf{u}_1, \dots, \mathbf{u}_n$  of the subspace  $W$ . Since, by definition,  $\langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0$  for  $i \neq j$ , while  $\langle \mathbf{u}_i, \mathbf{u}_i \rangle = \|\mathbf{u}_i\|^2 = 1$ , the associated Gram matrix is the identity matrix:  $K = I$ . Thus, in this situation, the system (5.24) reduces to simply  $\mathbf{x} = \mathbf{f}$ , with solution  $x_i^* = f_i = \langle \mathbf{u}_i, \mathbf{b} \rangle$ , and the closest point is given by

$$\mathbf{w}^* = x_1^* \mathbf{u}_1 + \dots + x_n^* \mathbf{u}_n \quad \text{where} \quad x_i^* = \langle \mathbf{b}, \mathbf{u}_i \rangle, \quad i = 1, \dots, n. \quad (5.32)$$

We have already seen this formula! According to Theorem 4.32,  $\mathbf{w}^*$  is the orthogonal projection of  $\mathbf{b}$  onto the subspace  $W$ . Thus, if we are supplied with an orthonormal basis of our subspace, we can easily compute the closest point using the orthogonal projection formula (5.32). If the basis is orthogonal, one can either normalize it or directly apply the equivalent orthogonal projection formula (4.42).

In this manner, we have established the key connection identifying the closest point in the subspace to a given vector with the orthogonal projection of that vector onto the subspace.

**Theorem 5.7.** Let  $W \subset V$  be a finite-dimensional subspace of an inner product space. Given a point  $\mathbf{b} \in V$ , the closest point  $\mathbf{w}^* \in W$  coincides with the orthogonal projection of  $\mathbf{b}$  onto  $W$ .

**Example 5.8.** Let  $\mathbb{R}^4$  be equipped with the ordinary Euclidean norm. Consider the three-dimensional subspace  $W \subset \mathbb{R}^4$  spanned by the orthogonal vectors  $\mathbf{v}_1 = (1, -1, 2, 0)^T$ ,  $\mathbf{v}_2 = (0, 2, 1, -2)^T$ ,  $\mathbf{v}_3 = (1, 1, 0, 1)^T$ . Given  $\mathbf{b} = (1, 2, 2, 1)^T$ , our task is to find the closest point  $\mathbf{w}^* \in W$ . Since the spanning vectors are orthogonal (but not orthonormal), we can use the orthogonal projection formula (4.42) to find  $\mathbf{w}^* = x_1 \mathbf{v}_1 + x_2 \mathbf{v}_2 + x_3 \mathbf{v}_3$ , with

$$x_1 = \frac{\langle \mathbf{b}, \mathbf{v}_1 \rangle}{\|\mathbf{v}_1\|^2} = \frac{3}{6} = \frac{1}{2}, \quad x_2 = \frac{\langle \mathbf{b}, \mathbf{v}_2 \rangle}{\|\mathbf{v}_2\|^2} = \frac{4}{9}, \quad x_3 = \frac{\langle \mathbf{b}, \mathbf{v}_3 \rangle}{\|\mathbf{v}_3\|^2} = \frac{4}{3}.$$

Thus, the closest point to  $\mathbf{b}$  in the given subspace is

$$\mathbf{w}^* = \frac{1}{2}\mathbf{v}_1 + \frac{4}{9}\mathbf{v}_2 + \frac{4}{3}\mathbf{v}_3 = \left( \frac{11}{6}, \frac{31}{18}, \frac{13}{9}, \frac{4}{9} \right)^T.$$

We further note that, in accordance with the orthogonal projection property, the vector

$$\mathbf{z} = \mathbf{b} - \mathbf{w}^* = \left( -\frac{5}{6}, \frac{5}{18}, \frac{5}{9}, \frac{5}{9} \right)^T$$

is orthogonal to  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  and hence to the entire subspace.

Even when we only know a non-orthogonal basis for the subspace, it may still be a good strategy to first apply the Gram–Schmidt process in order to replace it by an orthonormal or orthogonal basis, and then apply the relevant orthogonal projection formula to calculate the closest point. Not only does this simplify the final computation, it can often ameliorate the numerical inaccuracies associated with ill-conditioning that can afflict the direct solution to the system (5.24). The following example illustrates this alternative procedure.

**Example 5.9.** Let us return to the problem, solved in Example 5.6, of finding the closest point in the plane  $W$  spanned by  $\mathbf{w}_1 = (1, 2, -1)^T$ ,  $\mathbf{w}_2 = (2, -3, -1)^T$  to the point  $\mathbf{b} = (1, 0, 0)^T$ . We proceed by first using the Gram–Schmidt process to compute an orthogonal basis

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 = \begin{pmatrix} \frac{5}{2} \\ -2 \\ -\frac{3}{2} \end{pmatrix},$$

for our subspace. As a result, we can use the orthogonal projection formula (4.42) to produce the closest point

$$\mathbf{w}^* = \frac{\mathbf{b} \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1 + \frac{\mathbf{b} \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} \mathbf{v}_2 = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{15} \\ -\frac{7}{15} \end{pmatrix},$$

reconfirming our earlier result.

## Exercises

*Note:* Unless otherwise indicated, “distance” refers to the Euclidean norm.

5.3.1. Find the closest point in the plane spanned by  $(1, 2, -1)^T, (0, -1, 3)^T$  to the point  $(1, 1, 1)^T$ . What is the distance between the point and the plane?

5.3.2. Redo Exercise 5.3.1 using

(a) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1w_1 + 4v_2w_2 + 3v_3w_3$ ; (b) the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w}$  based on the positive definite matrix  $C = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ .

5.3.3. Find the point in the plane  $x + 2y - z = 0$  that is closest to  $(0, 0, 1)^T$ .

5.3.4. Let  $\mathbf{b} = (3, 1, 2, 1)^T$ . Find the closest point and the distance from  $\mathbf{b}$  to the following subspaces: (a) the line in the direction  $(1, 1, 1, 1)^T$ ; (b) the plane spanned by  $(1, 1, 0, 0)^T$  and  $(0, 0, 1, 1)^T$ ; (c) the hyperplane spanned by  $(1, 0, 0, 0)^T, (0, 1, 0, 0)^T, (0, 0, 1, 0)^T$ ; (d) the hyperplane defined by the equation  $x + y + z + w = 0$ .

5.3.5. Find the closest point and the distance from  $\mathbf{b} = (1, 1, 2, -2)^T$  to the subspace spanned by  $(1, 2, -1, 0)^T, (0, 1, -2, -1)^T, (1, 0, 3, 2)^T$ .

5.3.6. Redo Exercises 5.3.4 and 5.3.5 using

(i) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \frac{1}{2}v_1w_1 + v_2w_2 + \frac{1}{2}v_3w_3 + v_4w_4$ ;

(ii) the inner product based on the positive definite matrix  $C = \begin{pmatrix} 4 & -1 & 1 & 0 \\ -1 & 4 & -1 & 1 \\ 1 & -1 & 4 & -1 \\ 0 & 1 & -1 & 4 \end{pmatrix}$ .

5.3.7. Find the vector  $\mathbf{w}^* \in \text{span} \{ (0, 0, 1, 1), (2, 1, 1, 1) \}$  that minimizes  $\|\mathbf{w} - (0, 3, 1, 2)\|$ .

5.3.8. (a) Find the distance from the point  $\mathbf{b} = (1, 2, -1)^T$  to the plane  $x - 2y + z = 0$ .

(b) Find the distance to the plane  $x - 2y + z = 3$ .

*Hint:* Move the point and the plane so that the plane goes through the origin.

◇ 5.3.9. (a) Given a configuration of  $n$  points  $\mathbf{a}_1, \dots, \mathbf{a}_n$  in the plane, explain how to find the

point  $\mathbf{x} \in \mathbb{R}^2$  that minimizes the total squared distance  $\sum_{i=1}^n \|\mathbf{x} - \mathbf{a}_i\|^2$ . (b) Apply your

method when (i)  $\mathbf{a}_1 = (1, 3), \mathbf{a}_2 = (-2, 5)$ ; (ii)  $\mathbf{a}_1 = (0, 0), \mathbf{a}_2 = (0, 1), \mathbf{a}_3 = (1, 0)$ ;

(iii)  $\mathbf{a}_1 = (0, 0), \mathbf{a}_2 = (0, 2), \mathbf{a}_3 = (1, 2), \mathbf{a}_4 = (-2, -1)$ .

5.3.10. Answer Exercise 5.3.9 when distance is measured in (a) the weighted norm

$\|\mathbf{x}\| = \sqrt{2x_1^2 + 3x_2^2}$ ; (b) the norm based on the positive definite matrix  $\begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$ .

5.3.11. Explain why the quantity inside the square root in (5.25) is always non-negative.

5.3.12. Find the closest point to the vector  $\mathbf{b} = (1, 0, 2)^T$  belonging the two-dimensional subspace spanned by the orthogonal vectors  $\mathbf{v}_1 = (1, -1, 1)^T, \mathbf{v}_2 = (-1, 1, 2)^T$ .

5.3.13. Let  $\mathbf{b} = (0, 3, 1, 2)^T$ . Find the vector  $\mathbf{w}^* \in \text{span} \{ (0, 0, 1, 1)^T, (2, 1, 1, -1)^T \}$  such that  $\|\mathbf{w}^* - \mathbf{b}\|$  is minimized.

5.3.14. Find the closest point to  $\mathbf{b} = (1, 2, -1, 3)^T$  in the subspace  $W = \text{span} \{ (1, 0, 2, 1)^T, (1, 1, 0, -1)^T, (2, 0, 1, -1)^T \}$  by first constructing an orthogonal basis of  $W$  and then applying the orthogonal projection formula (4.42).

5.3.15. Repeat Exercise 5.3.14 using the weighted norm  $\|\mathbf{v}\| = v_1^2 + 2v_2^2 + v_3^2 + 3v_4^2$ .

◇ 5.3.16. Justify the formulas in (5.29).

## 5.4 Least Squares

As we first observed in Section 5.1, the solution to the closest point problem also solves the basic least squares minimization problem. Let us first officially define the notion of a (classical) least squares solution to a linear system of equations.

**Definition 5.10.** A *least squares solution* to a linear system of equations

$$A\mathbf{x} = \mathbf{b} \tag{5.33}$$

is a vector  $\mathbf{x}^* \in \mathbb{R}^n$  that minimizes the squared Euclidean norm  $\|A\mathbf{x} - \mathbf{b}\|^2$ .

If the system (5.33) actually has a solution, then it is automatically the least squares

solution. The concept of least squares solution is new only when the system does not have a solution, i.e.,  $\mathbf{b}$  does not lie in the image of  $A$ . We also want the least squares solution to be unique. As with an ordinary solution, this happens if and only if  $\ker A = \{\mathbf{0}\}$ , or, equivalently, the columns of  $A$  are linearly independent, or, equivalently,  $\text{rank } A = n$ . Indeed, if  $\mathbf{z} \in \ker A$ , then  $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{z}$  also satisfies

$$\|A\tilde{\mathbf{x}} - \mathbf{b}\|^2 = \|A(\mathbf{x} + \mathbf{z}) - \mathbf{b}\|^2 = \|A\mathbf{x} - \mathbf{b}\|^2,$$

and hence is also a minimum. Thus, uniqueness requires  $\mathbf{z} = \mathbf{0}$ .

As before, to make the connection with the closest point problem, we identify the subspace  $W = \text{img } A \subset \mathbb{R}^m$  as the image or column space of the matrix  $A$ . If the columns of  $A$  are linearly independent, then they form a basis for the image  $W$ . Since every element of the image can be written as  $\mathbf{w} = A\mathbf{x}$ , minimizing  $\|A\mathbf{x} - \mathbf{b}\|^2$  is the same as minimizing the distance  $\|\mathbf{w} - \mathbf{b}\|$  between the point and the subspace. The solution  $\mathbf{x}^*$  to the quadratic minimization problem produces the closest point  $\mathbf{w}^* = A\mathbf{x}^*$  in  $W = \text{img } A$ , which is thus found using Theorem 5.5. In the Euclidean case, we therefore find the least squares solution by solving the normal equations given in (5.27).

**Theorem 5.11.** Assume that  $\ker A = \{\mathbf{0}\}$ . Then the least squares solution to the linear system  $A\mathbf{x} = \mathbf{b}$  under the Euclidean norm is the unique solution  $\mathbf{x}^*$  to the *normal equations*

$$(A^T A)\mathbf{x} = A^T \mathbf{b}, \quad \text{namely} \quad \mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{b}. \quad (5.34)$$

The *least squares error* is

$$\|A\mathbf{x}^* - \mathbf{b}\|^2 = \|\mathbf{b}\|^2 - \mathbf{f}^T \mathbf{x}^* = \|\mathbf{b}\|^2 - \mathbf{b}^T A (A^T A)^{-1} A^T \mathbf{b}. \quad (5.35)$$

Note that the normal equations (5.27) can be simply obtained by multiplying the original system  $A\mathbf{x} = \mathbf{b}$  on both sides by  $A^T$ . In particular, if  $A$  is square and invertible, then  $(A^T A)^{-1} = A^{-1}(A^T)^{-1}$ , and so the least squares solution formula (5.34) reduces to  $\mathbf{x} = A^{-1}\mathbf{b}$ , while the two terms under the square root in the error formula (5.35) cancel out, producing zero error. In the rectangular case — when inversion of  $A$  itself is *not* allowed — (5.34) gives a *new* formula for the solution to the linear system  $A\mathbf{x} = \mathbf{b}$  whenever  $\mathbf{b} \in \text{img } A$ . See also the discussion concerning the pseudoinverse of a matrix in Section 8.7 for an alternative approach.

**Example 5.12.** Consider the linear system

$$\begin{aligned} x_1 + 2x_2 &= 1, \\ 3x_1 - x_2 + x_3 &= 0, \\ -x_1 + 2x_2 + x_3 &= -1, \\ x_1 - x_2 - 2x_3 &= 2, \\ 2x_1 + x_2 - x_3 &= 2, \end{aligned}$$

consisting of 5 equations in 3 unknowns. The coefficient matrix and right-hand side are

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & -1 & -2 \\ 2 & 1 & -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \\ 2 \end{pmatrix}.$$

A direct application of Gaussian Elimination shows that  $\mathbf{b} \notin \text{img } A$ , and so the system is incompatible — it has no solution. Of course, to apply the least squares method, we are not required to check this in advance. If the system has a solution, it is the least squares solution too, and the least squares method will find it.

Let us find the least squares solution based on the Euclidean norm, so that  $C = I$  in Theorem 5.13. According to (5.26),

$$K = A^T A = \begin{pmatrix} 16 & -2 & -2 \\ -2 & 11 & 2 \\ -2 & 2 & 7 \end{pmatrix}, \quad \mathbf{f} = A^T \mathbf{b} = \begin{pmatrix} 8 \\ 0 \\ -7 \end{pmatrix}.$$

Solving the  $3 \times 3$  system of normal equations  $K \mathbf{x} = \mathbf{f}$  by Gaussian Elimination, we find

$$\mathbf{x}^* = K^{-1} \mathbf{f} \approx (.4119, .2482, -.9532)^T$$

to be the least squares solution to the system. The least squares error is

$$\| \mathbf{b} - A \mathbf{x}^* \|^2 \approx \| (-.0917, .0342, .1313, .0701, .0252)^T \|^2 \approx .03236,$$

which is reasonably small — indicating that the system is, roughly speaking, not too incompatible.

An alternative strategy is to begin by orthonormalizing the columns of  $A$  using Gram–Schmidt. We can then apply the orthogonal projection formula (4.41) to construct the same least squares solution. Details of the latter computation are left to the reader.

One can extend the basic least squares method by introducing a suitable weighted norm in the measurement of the error. Let  $C > 0$  be a positive definite matrix that governs the weighted norm  $\| \mathbf{v} \|^2 = \mathbf{v}^T C \mathbf{v}$ . In most applications,  $C = \text{diag}(c_1, \dots, c_m)$  is a diagonal matrix whose entries are the assigned weights of the individual coordinates, but the method works equally well for general norms defined by positive definite matrices. The off-diagonal entries of  $C$  can be used to weight cross-correlations between data values, although this extra freedom is rarely used in practice. The weighted least squares solution is thus obtained by solving the corresponding weighted normal equations (5.30), as follows.

**Theorem 5.13.** Suppose  $A$  is an  $m \times n$  matrix such that  $\ker A = \{\mathbf{0}\}$ , and suppose  $C > 0$  is any positive definite  $m \times m$  matrix specifying the weighted norm  $\| \mathbf{v} \|^2 = \mathbf{v}^T C \mathbf{v}$ . Then the least squares solution to the linear system  $A \mathbf{x} = \mathbf{b}$  that minimizes the weighted squared error  $\| A \mathbf{x} - \mathbf{b} \|^2$  is the unique solution  $\mathbf{x}^*$  to the *weighted normal equations*

$$A^T C A \mathbf{x}^* = A^T C \mathbf{b}, \quad \text{so that} \quad \mathbf{x}^* = (A^T C A)^{-1} A^T C \mathbf{b}. \quad (5.36)$$

The *weighted least squares error* is

$$\| A \mathbf{x}^* - \mathbf{b} \|^2 = \| \mathbf{b} \|^2 - \mathbf{f}^T \mathbf{x}^* = \| \mathbf{b} \|^2 - \mathbf{b}^T C A (A^T A)^{-1} A^T C \mathbf{b}. \quad (5.37)$$

## Exercises

*Note:* Unless otherwise indicated, use the Euclidean norm to measure the least squares error.

5.4.1. Find the least squares solution to the linear system  $A \mathbf{x} = \mathbf{b}$  when

$$(a) A = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad (b) A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 3 & 5 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix}, \quad (c) A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & -2 & 0 \\ 3 & -1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

5.4.2. Find the least squares solutions to the following linear systems:

- (a)  $x + 2y = 1, 3x - y = 0, -x + 2y = 3$ ,
- (b)  $4x - 2y = 1, 2x + 3y = -4, x - 2y = -1, 2x + 2y = 2$ ,
- (c)  $2u + v - 2w = 1, 3u - 2w = 0, u - v + 3w = 2$ ,
- (d)  $x - z = -1, 2x - y + 3z = 1, y - 3z = 0, -5x + 2y + z = 3$ ,
- (e)  $x_1 + x_2 = 2, x_2 + x_4 = 1, x_1 + x_3 = 0, x_3 - x_4 = 1, x_1 - x_4 = 2$ .

5.4.3. Let  $A = \begin{pmatrix} 3 & -3 & 1 \\ 2 & 4 & 1 \\ 1 & 2 & 1 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 6 \\ 5 \\ 4 \end{pmatrix}$ . Prove, using Gaussian Elimination, that the linear

system  $A\mathbf{x} = \mathbf{b}$  has a unique solution. Show that the least squares solution (5.34) is the same. Explain why this is necessarily the case.

5.4.4. Find the least squares solution to the linear system  $A\mathbf{x} = \mathbf{b}$  when

$$(a) A = \begin{pmatrix} 2 & 3 \\ 4 & -2 \\ 1 & 5 \\ 2 & 0 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 3 \end{pmatrix}, \quad (b) A = \begin{pmatrix} 2 & 1 & 4 \\ 1 & -2 & 1 \\ 1 & 0 & -3 \\ 5 & 2 & -2 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

5.4.5. Given  $A = \begin{pmatrix} 1 & 2 & -1 \\ 0 & -2 & 3 \\ 1 & 5 & -1 \\ -3 & 1 & 1 \end{pmatrix}$  and  $\mathbf{b} = \begin{pmatrix} 0 \\ 5 \\ 6 \\ 8 \end{pmatrix}$ , find the least squares solution to the system  $A\mathbf{x} = \mathbf{b}$ . What is the error? Interpret your result.

5.4.6. Find the least squares solution to the linear systems in Exercise 5.4.1 under the weighted norm  $\|\mathbf{x}\|^2 = x_1^2 + 2x_2^2 + 3x_3^2$ .

◇ 5.4.7. Let  $A$  be an  $m \times n$  matrix with  $\ker A = \{\mathbf{0}\}$ . Suppose that we use the Gram–Schmidt algorithm to factor  $A = QR$  as in Exercise 4.3.32. Prove that the least squares solution to the linear system  $A\mathbf{x} = \mathbf{b}$  is found by solving the triangular system  $R\mathbf{x} = Q^T\mathbf{b}$  by Back Substitution.

5.4.8. Apply the method in Exercise 5.4.7 to find the least squares solutions to the systems in Exercise 5.4.2.

◇ 5.4.9. (a) Find a formula for the least squares error (5.35) in terms of an orthonormal basis of the subspace. (b) Generalize your formula to the case of an orthogonal basis.

5.4.10. Find the least squares solutions to the following linear systems. *Hint:* Check

orthogonality of the columns of the coefficient matrix. (a)  $\begin{pmatrix} 1 & -1 \\ 2 & 2 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$ ,

(b)  $\begin{pmatrix} 3 & -1 \\ 0 & 2 \\ -2 & 1 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ -1 \\ 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & -1 & -1 \\ 1 & 3 & 2 \\ -2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 7 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}$ .

◇ 5.4.11. Suppose we are interested in solving a linear system  $A\mathbf{x} = \mathbf{b}$  by the method of least squares when the coefficient matrix  $A$  has linearly dependent columns. Let  $K\mathbf{x} = \mathbf{f}$ , where  $K = A^TCA$ ,  $\mathbf{f} = A^TC\mathbf{b}$ , be the corresponding normal equations. (a) Prove that  $\mathbf{f} \in \text{img } K$ , and so the normal equations have a solution. *Hint:* Use Exercise 3.4.32. (b) Prove that every solution to the normal equations minimizes the least squares error, and hence qualifies as a least squares solution to the original system. (c) Explain why the least squares solution is not unique.

◇ 5.4.12. Which is the more efficient algorithm: direct least squares based on solving the normal equations by Gaussian Elimination, or using Gram–Schmidt orthonormalization and then solving the resulting triangular system by Back Substitution as in Exercise 5.4.7? Justify your answer.

5.4.13. A group of students knows that the least squares solution to  $A \mathbf{x} = \mathbf{b}$  can be identified with the closest point on the subspace  $\text{img } A$  spanned by the columns of the coefficient matrix. Therefore, they try to find the solution by first orthonormalizing the columns using Gram–Schmidt, and then finding the least squares coefficients by the orthonormal basis formula (4.41). To their surprise, they do not get the same solution! Can you explain the source of their difficulty? How can you use their solution to obtain the proper least squares solution  $\mathbf{x}$ ? Check your algorithm with the system that we treated in Example 5.12.

## 5.5 Data Fitting and Interpolation

One of the most important applications of the least squares minimization process is to the fitting of data points. Suppose we are running an experiment in which we measure a certain time-dependent physical quantity. At time  $t_i$  we make the measurement  $y_i$ , and thereby obtain a set of, say,  $m$  data points

$$(t_1, y_1), \quad (t_2, y_2), \quad \dots \quad (t_m, y_m). \quad (5.38)$$

Suppose our theory indicates that all the data points are supposed to lie on a single line

$$y = \alpha + \beta t, \quad (5.39)$$

whose precise form — meaning its coefficients  $\alpha, \beta$  — is to be determined. For example, a police car is interested in clocking the speed of a vehicle by using measurements of its relative distance at several times. Assuming that the vehicle is traveling at constant speed, its position at time  $t$  will have the linear form (5.39), with  $\beta$ , the velocity, and  $\alpha$ , the initial position, to be determined. The amount by which  $\beta$  exceeds the speed limit will determine whether the police decide to give chase. Experimental error will almost inevitably make this measurement impossible to achieve exactly, and so the problem is to find the straight line (5.39) that “best fits” the measured data and then use its slope to estimate the vehicle’s velocity.

At the time  $t = t_i$ , the *error* between the measured value  $y_i$  and the sample value predicted by the function (5.39) is

$$e_i = y_i - (\alpha + \beta t_i), \quad i = 1, \dots, m.$$

We can write this system of equations in the compact vectorial form

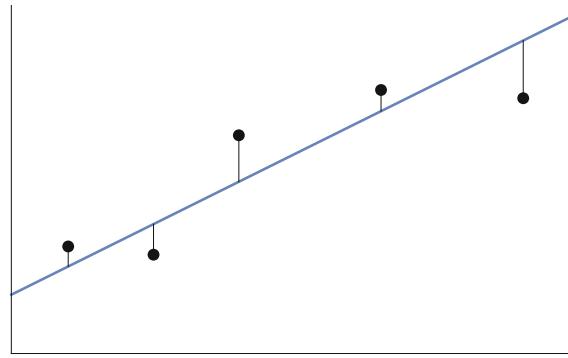
$$\mathbf{e} = \mathbf{y} - A \mathbf{x},$$

where

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \text{while} \quad A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \quad (5.40)$$

We call  $\mathbf{e} \in \mathbb{R}^m$  the *error vector* and  $\mathbf{y} \in \mathbb{R}^m$  the *data vector*. The  $m \times 2$  matrix  $A$  is prescribed by the sample times. The coefficients  $\alpha, \beta$  of our desired function (5.39) are the unknowns, forming the entries of the column vector  $\mathbf{x} \in \mathbb{R}^2$ .

If we could fit the data exactly, so  $y_i = \alpha + \beta t_i$  for all  $i$ , then each error would vanish,  $e_i = 0$ , and we could solve the linear system  $A \mathbf{x} = \mathbf{y}$  for the coefficients  $\alpha, \beta$ . In the language of linear algebra, the data points all lie on a straight line if and only if  $\mathbf{y} \in \text{img } A$ .



**Figure 5.5.** Least Squares Approximation of Data by a Straight Line.

If the data points are not collinear, then we seek the straight line that minimizes the total *squared error*:

$$\text{Squared Error} = \|\mathbf{e}\|^2 = e_1^2 + \cdots + e_m^2,$$

which coincides with the squared Euclidean norm of the error vector. Pictorially, referring to Figure 5.5, the errors are the vertical distances from the points to the line, and we are seeking to minimize the sum of the squares of the individual errors<sup>†</sup>, hence the term *least squares*. In other words, we are looking for the coefficient vector  $\mathbf{x} = (\alpha, \beta)^T$  that minimizes the Euclidean norm of the error vector

$$\|\mathbf{e}\| = \|A\mathbf{x} - \mathbf{y}\|. \quad (5.41)$$

Thus, we have a manifestation of the problem of characterizing the least squares solution to the linear system  $A\mathbf{x} = \mathbf{y}$ .

Theorem 5.11 prescribes the solution to this least squares minimization problem. We form the normal equations

$$(A^T A)\mathbf{x} = A^T \mathbf{y}, \quad \text{with solution} \quad \mathbf{x}^* = (A^T A)^{-1} A^T \mathbf{y}. \quad (5.42)$$

Invertibility of the Gram matrix  $K = A^T A$  relies on the assumption that the matrix  $A$  has linearly independent columns. For the particular matrix in (5.40), linear independence of its two columns requires that not all the  $t_i$ 's be equal, i.e., we must measure the data at at least two distinct times. Note that this restriction does not preclude measuring some of the data at the same time, e.g., by repeating the experiment. However, choosing *all* the  $t_i$ 's to be the same is a silly data fitting problem. (Why?)

<sup>†</sup> This choice of minimization may strike the reader as a little odd. Why not just minimize the sum of the absolute value of the errors, i.e., the 1 norm  $\|\mathbf{e}\|_1 = |e_1| + \cdots + |e_n|$  of the error vector, or minimize the maximal error, i.e., the  $\infty$  norm  $\|\mathbf{e}\|_\infty = \max\{|e_1|, \dots, |e_n|\}$ ? The answer is that, although each of these alternative minimization criteria is interesting and potentially useful, they all lead to *nonlinear* minimization problems, and so are much harder to solve! The least squares minimization problem can be solved by linear algebra, and so, purely on the grounds of simplicity, is the method of choice in most applications. Moreover, as always, one needs to fully understand the linear problem before diving into more treacherous nonlinear waters. Or, even better, why minimize the vertical distance to the line? The shortest distance from each data point to the line, as measured along the perpendicular and explicitly computed in Exercise 5.1.8, might strike you as a better measure of error. To solve the latter problem, see Section 8.8 on Principal Component Analysis, particularly Exercise 8.8.11.

Under this assumption, we then compute

$$A^T A = \begin{pmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_m \end{pmatrix} \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} = \begin{pmatrix} m & \sum t_i \\ \sum t_i & \sum (t_i)^2 \end{pmatrix} = m \begin{pmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{pmatrix}, \quad (5.43)$$

$$A^T \mathbf{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_m \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum t_i y_i \end{pmatrix} = m \begin{pmatrix} \bar{y} \\ \bar{t} \bar{y} \end{pmatrix},$$

where the overbars, namely

$$\bar{t} = \frac{1}{m} \sum_{i=1}^m t_i, \quad \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i, \quad \bar{t}^2 = \frac{1}{m} \sum_{i=1}^m t_i^2, \quad \bar{t} \bar{y} = \frac{1}{m} \sum_{i=1}^m t_i y_i, \quad (5.44)$$

denote the *average* sample values of the indicated variables.

**Warning.** The average of a product is *not* equal to the product of the averages! In particular,  $\bar{t}^2 \neq (\bar{t})^2$ ,  $\bar{t} \bar{y} \neq \bar{t} \bar{y}$ .

Substituting (5.43) into the normal equations (5.42), and canceling the common factor of  $m$ , we find that we have only to solve the pair of linear equations

$$\alpha + \beta \bar{t} = \bar{y}, \quad \alpha \bar{t} + \beta \bar{t}^2 = \bar{t} \bar{y},$$

for the coefficients:

$$\alpha = \bar{y} - \beta \bar{t}, \quad \beta = \frac{\bar{t} \bar{y} - \bar{t} \bar{y}}{\bar{t}^2 - (\bar{t})^2} = \frac{\sum (t_i - \bar{t}) y_i}{\sum (t_i - \bar{t})^2}. \quad (5.45)$$

Therefore, the best (in the least squares sense) straight line that fits the given data is

$$y = \beta (t - \bar{t}) + \bar{y}, \quad (5.46)$$

where the line's slope  $\beta$  is given in (5.45).

More generally, one may wish to assign different weights to the measurement errors. Suppose some of the data are known to be more reliable or more significant than others. For example, measurements at an earlier time may be more accurate, or more critical to the data fitting problem, than later measurements. In that situation, we should penalize any errors in the earlier measurements and downplay errors in the later data.

In general, this requires the introduction of a positive weight  $c_i > 0$  associated with each data point  $(t_i, y_i)$ ; the larger the weight, the more vital the error. For a straight line approximation  $y = \alpha + \beta t$ , the *weighted squared error* is defined as

$$\text{Weighted Squared Error} = \sum_{i=1}^m c_i e_i^2 = \mathbf{e}^T C \mathbf{e} = \|\mathbf{e}\|^2,$$

where  $C = \text{diag}(c_1, \dots, c_m) > 0$  is the positive definite diagonal *weight matrix*, while  $\|\mathbf{e}\|$  denotes the associated weighted norm of the error vector  $\mathbf{e} = \mathbf{y} - A \mathbf{x}$ . One then applies the weighted normal equations (5.30) to effect the solution to the problem.

**Example 5.14.** Suppose the data points are given by the table

$t_i$	0	1	3	6
$y_i$	2	3	7	12

To find the least squares line, we construct

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix}, \quad A^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix}.$$

Therefore

$$A^T A = \begin{pmatrix} 4 & 10 \\ 10 & 46 \end{pmatrix}, \quad A^T \mathbf{y} = \begin{pmatrix} 24 \\ 96 \end{pmatrix}.$$

The normal equations (5.42) reduce to

$$4\alpha + 10\beta = 24, \quad 10\alpha + 46\beta = 96, \quad \text{so} \quad \alpha = \frac{12}{7}, \quad \beta = \frac{12}{7}.$$

Therefore, the best least squares fit to the data is the straight line

$$y = \frac{12}{7} + \frac{12}{7}t \approx 1.71429 + 1.71429t.$$

Alternatively, one can compute this formula directly from (5.45–46).

Now, suppose we assign different weights to the preceding data points, e.g.,  $c_1 = 3$ ,  $c_2 = 2$ ,  $c_3 = \frac{1}{2}$ ,  $c_4 = \frac{1}{4}$ . Thus, errors in the first two data values are assigned higher significance than those in the latter two. To find the weighted least squares line that best fits the data, we compute

$$A^T C A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} \frac{23}{4} & 5 \\ 5 & \frac{31}{2} \end{pmatrix},$$

$$A^T C \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix} = \begin{pmatrix} \frac{37}{2} \\ \frac{69}{2} \end{pmatrix}.$$

Thus, the weighted normal equations (5.30) reduce to

$$\frac{23}{4}\alpha + 5\beta = \frac{37}{2}, \quad 5\alpha + \frac{31}{2}\beta = \frac{69}{2}, \quad \text{so} \quad \alpha = 1.7817, \quad \beta = 1.6511.$$

Therefore, the least squares fit to the data under the given weights is

$$y = 1.7817 + 1.6511t.$$

**Example 5.15.** Suppose we are given a sample of an unknown radioactive isotope. At several times  $t_i$ , we measure the amount  $m_i$  of radioactive material remaining in the sample. The problem is to determine the initial amount of material along with the isotope's half-life. If the measurements were exact, we would have  $m(t) = m_0 e^{\beta t}$ , where  $m_0 = m(0)$  is the initial mass, and  $\beta < 0$  the decay rate. The half-life is given by  $t^* = \beta^{-1} \log 2$ ; see Example 8.1 for additional details.

As it stands, this is *not* a linear least squares problem. But it can be easily converted to the proper form by taking logarithms:

$$y(t) = \log m(t) = \log m_0 + \beta t = \alpha + \beta t \quad \text{where} \quad \alpha = \log m_0.$$

We can thus do a linear least squares fit on the logarithms  $y_i = \log m_i$  of the radioactive mass data at the measurement times  $t_i$  to determine the best values for  $\alpha$  and  $\beta$ .

## Exercises

5.5.1. Find the straight line  $y = \alpha + \beta t$  that best fits the following data in the least squares

$t_i$	-2	0	1	3
$y_i$	0	1	2	5

 $|       |   |   |    |    |    | |-------|---|---|----|----|----| | $t_i$ | 1 | 2 | 3  | 4  | 5  | | $y_i$ | 1 | 0 | -2 | -3 | -3 |$ |



sense: (a)	$   t_i-2013  y_i0125  $	$t_i$	-2	0	1	3	$y_i$	0	1	2	5	$   t_i12345  y_i10-2-3-3  $	$t_i$	1	2	3	4	5	$y_i$	1	0	-2	-3	-3
$t_i$	-2	0	1	3																				
$y_i$	0	1	2	5																				
$t_i$	1	2	3	4	5																			
$y_i$	1	0	-2	-3	-3																			
(c)	$   t_i-2-1012  y_i-5-3-203  $	$t_i$	-2	-1	0	1	2	$y_i$	-5	-3	-2	0	3											
$t_i$	-2	-1	0	1	2																			
$y_i$	-5	-3	-2	0	3																			

5.5.2. The proprietor of an internet travel company compiled the following data relating the annual profit of the firm to its annual advertising expenditure (both measured in thousands of dollars):

Annual advertising expenditure	12	14	17	21	26	30
Annual profit	60	70	90	100	100	120

- (a) Determine the equation of the least squares line. (b) Plot the data and the least squares line. (c) Estimate the profit when the annual advertising budget is \$50,000. (d) What about a \$100,000 budget?

5.5.3. The median price (in thousands of dollars) of existing homes in a certain metropolitan area from 1989 to 1999 was:

year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
price	86.4	89.8	92.8	96.0	99.6	103.1	106.3	109.5	113.3	120.0	129.5

- (a) Find an equation of the least squares line for these data. (b) Estimate the median price of a house in the year 2005, and the year 2010, assuming that the trend continues.

5.5.4. A 20-pound turkey that is at the room temperature of  $72^\circ$  is placed in the oven at 1:00 pm. The temperature of the turkey is observed in 20 minute intervals to be  $79^\circ$ ,  $88^\circ$ , and  $96^\circ$ . A turkey is cooked when its temperature reaches  $165^\circ$ . How much longer do you need to wait until the turkey is done?

5.5.5. The amount of waste (in millions of tons a day) generated in a certain city from 1960 to 1995 was

year	1960	1965	1970	1975	1980	1985	1990	1995
amount	86	99.8	115.8	125	132.6	143.1	156.3	169.5

- (a) Find the equation for the least squares line that best fits these data. (b) Use the result to estimate the amount of waste in the year 2000, and in the year 2005. (c) Redo your calculations using an exponential growth model  $y = ce^{\alpha t}$ . (d) Which model do you think most accurately reflects the data? Why?

5.5.6. The amount of radium-224 in a sample was measured at the indicated times.

time in days	0	1	2	3	4	5	6	7
mg	100	82.7	68.3	56.5	46.7	38.6	31.9	26.4

- (a) Estimate how much radium will be left after 10 days.  
 (b) If the sample is considered to be safe when the amount of radium is less than .01 mg, estimate how long the sample needs to be stored before it can be safely disposed of.

5.5.7. The following table gives the population of the United States for the years 1900-2000.

year	1900	1920	1940	1960	1980	2000
population – in millions	76	106	132	181	227	282

- (a) Use an exponential growth model of the form  $y = ce^{at}$  to predict the population in 2020, 2050, and 3000. (b) The actual population for the year 2020 has recently been estimated to be 334 million. How does this affect your predictions for 2050 and 3000?

5.5.8. Find the best linear least squares fit of the following data using the indicated weights:

	$t_i$	1	2	3	4
(a)	$y_i$	.2	.4	.7	1.2
	$c_i$	1	2	3	4

	$t_i$	0	1	3	6
(b)	$y_i$	2	3	7	12
	$c_i$	4	3	2	1

	$t_i$	-2	-1	0	1	2
(c)	$y_i$	-5	-3	-2	0	3
	$c_i$	2	1	.5	1	2

	$t_i$	1	2	3	4	5
(d)	$y_i$	2	1.3	1.1	.8	.2
	$c_i$	5	4	3	2	1

5.5.9. For the data points

	$x$	1	1	2	2	3	3
	$y$	1	2	1	2	2	4
	$z$	3	6	11	-2	0	3

- (a) determine the best plane  $z = a + bx + cy$  that best fits the data in the least squares sense; (b) how would you answer the question in part (a) if the plane were constrained to go through the point  $x = 2, y = 2, z = 0$ ?

5.5.10. For the data points in Exercise 5.5.9, determine the plane  $z = \alpha + \beta x + \gamma y$  that fits the data in the least squares sense when the errors are weighted according to the reciprocal of the distance of the point  $(x_i, y_i, z_i)$  from the origin.

◊ 5.5.11. Show, by constructing explicit examples, that  $\bar{t^2} \neq (\bar{t})^2$  and  $\bar{ty} \neq \bar{t}\bar{y}$ . Can you find any data for which either equality is valid?

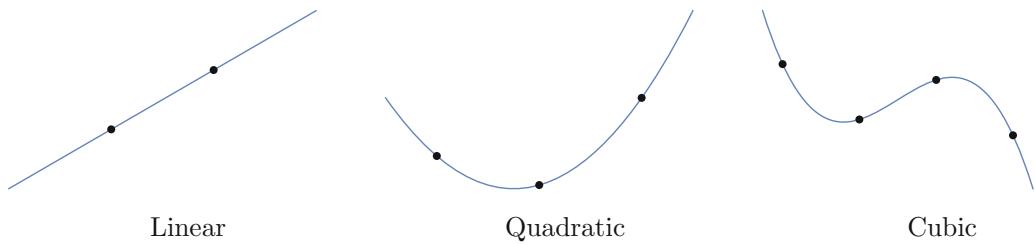
◊ 5.5.12. Given points  $t_1, \dots, t_m$ , prove  $\bar{t^2} - (\bar{t})^2 = \frac{1}{m} \sum_{i=1}^m (t_i - \bar{t})^2$ , thereby justifying (5.45).

## Polynomial Approximation and Interpolation

The basic least squares philosophy has a variety of different extensions, all interesting and all useful. First, we can replace the straight line (5.39) by a parabola defined by a quadratic function

$$y = \alpha + \beta t + \gamma t^2. \quad (5.47)$$

For example, Newton's theory of gravitation says that (in the absence of air resistance) a falling object obeys the parabolic law (5.47), where  $\alpha = h_0$  is the initial height,  $\beta = v_0$  is the initial velocity, and  $\gamma = -\frac{1}{2}g$  is minus one half the gravitational constant. Suppose we observe a falling body on a new planet, and measure its height  $y_i$  at times  $t_i$ . Then we



**Figure 5.6.** Interpolating Polynomials.

can approximate its initial height, initial velocity, and gravitational acceleration by finding the parabola (5.47) that best fits the data. Again, we characterize the least squares fit by minimizing the sum of the squares of the individual errors  $e_i = y_i - y(t_i)$ .

The method can evidently be extended to a completely general polynomial function

$$y(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n \quad (5.48)$$

of degree  $n$ . The total squared error between the data and the sample values of the function is equal to

$$\|\mathbf{e}\|^2 = \sum_{i=1}^m [y_i - y(t_i)]^2 = \|\mathbf{y} - A\mathbf{x}\|^2, \quad (5.49)$$

where

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}. \quad (5.50)$$

The  $m \times (n+1)$  coefficient matrix is known as a *Vandermonde matrix*, named after the eighteenth-century French mathematician, scientist, and musician/musicologist Alexandre-Théophile Vandermonde — despite the fact that it appears nowhere in his four mathematical papers! In particular, if  $m = n+1$ , then  $A$  is square, and so, assuming invertibility, we can solve  $A\mathbf{x} = \mathbf{y}$  exactly. In other words, there is no error, and the solution is an *interpolating polynomial*, meaning that it fits the data exactly. A proof of the following result can be found at the end of this section.

**Lemma 5.16.** If  $t_1, \dots, t_{n+1} \in \mathbb{R}$  are distinct, so  $t_i \neq t_j$  for  $i \neq j$ , then the  $(n+1) \times (n+1)$  Vandermonde interpolation matrix (5.50) is nonsingular.

This result implies the basic existence theorem for interpolating polynomials.

**Theorem 5.17.** Let  $t_1, \dots, t_{n+1}$  be distinct sample points. Then, for any prescribed data  $y_1, \dots, y_{n+1}$ , there exists a *unique* interpolating polynomial  $y(t)$  of degree  $\leq n$  that has the prescribed sample values:  $y(t_i) = y_i$  for all  $i = 1, \dots, n+1$ .

Thus, in particular, two points will determine a unique interpolating line, three points a unique interpolating parabola, four points an interpolating cubic, and so on, as sketched in Figure 5.6.

The basic ideas of interpolation and least squares fitting of data can be applied to approximate complicated mathematical functions by much simpler polynomials. Such approximation schemes are used in all numerical computations. Your computer or calculator

is only able to add, subtract, multiply, and divide. Thus, when you ask it to compute  $\sqrt{t}$  or  $e^t$  or  $\cos t$  or any other non-rational function, the program must rely on an approximation scheme based on polynomials<sup>†</sup>. In the “dark ages” before electronic computers<sup>‡</sup>, one would consult precomputed tables of values of the function at particular data points. If one needed a value at a non-tabulated point, then some form of polynomial interpolation would be used to approximate the intermediate value.

**Example 5.18.** Suppose that we would like to compute reasonably accurate values for the exponential function  $e^t$  for values of  $t$  lying in the interval  $0 \leq t \leq 1$  by approximating it by a quadratic polynomial

$$p(t) = \alpha + \beta t + \gamma t^2. \quad (5.51)$$

If we choose 3 points, say  $t_1 = 0$ ,  $t_2 = .5$ ,  $t_3 = 1$ , then there is a unique quadratic polynomial (5.51) that interpolates  $e^t$  at the data points, i.e.,  $p(t_i) = e^{t_i}$  for  $i = 1, 2, 3$ . In this case, the coefficient matrix (5.50), namely  $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .5 & .25 \\ 1 & 1 & 1 \end{pmatrix}$ , is nonsingular, so we can exactly solve the interpolation equations

$$A \mathbf{x} = \mathbf{y}, \quad \text{where} \quad \mathbf{y} = \begin{pmatrix} e^{t_1} \\ e^{t_2} \\ e^{t_3} \end{pmatrix} = \begin{pmatrix} 1. \\ 1.64872 \\ 2.71828 \end{pmatrix}$$

is the data vector, which we assume we already know. The solution

$$\mathbf{x} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 1. \\ .876603 \\ .841679 \end{pmatrix}$$

yields the interpolating polynomial

$$p(t) = 1 + .876603t + .841679t^2. \quad (5.52)$$

It is the unique quadratic polynomial that agrees with  $e^t$  at the three specified data points. See [Figure 5.7](#) for a comparison of the graphs; the first graph shows  $e^t$ , the second  $p(t)$ , and the third lays the two graphs on top of each other. Even with such a primitive interpolation scheme, the two functions are quite close. The maximum error, or  $L^\infty$  norm, of the difference is

$$\|e^t - p(t)\|_\infty = \max \{ |e^t - p(t)| \mid 0 \leq t \leq 1 \} \approx .01442,$$

with the largest deviation occurring at  $t \approx .796$ .

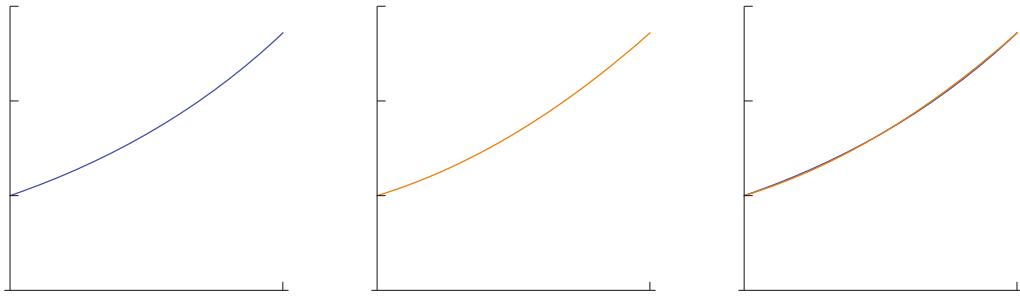
There is, in fact, an explicit formula for the interpolating polynomial that is named after the influential eighteenth century Italian–French mathematician Joseph-Louis Lagrange. It relies on the basic superposition principle for solving inhomogeneous systems that we found in [Theorem 2.44](#). Specifically, suppose we know the solutions  $\mathbf{x}_1, \dots, \mathbf{x}_{n+1}$  to the particular interpolation systems

$$A \mathbf{x}_k = \mathbf{e}_k, \quad k = 1, \dots, n+1, \quad (5.53)$$

---

<sup>†</sup> Actually, since division also is possible, one could also allow interpolation and approximation by rational functions, a subject known as *Padé approximation theory*, [3].

<sup>‡</sup> Back then, the word “computer” referred to a human who computed, mostly female and including the first author’s mother, Grace Olver.



**Figure 5.7.** Quadratic Interpolating Polynomial for  $e^t$ .

where  $\mathbf{e}_1, \dots, \mathbf{e}_{n+1}$  are the standard basis vectors of  $\mathbb{R}^{n+1}$ . Then the solution to

$$A \mathbf{x} = \mathbf{y} = y_1 \mathbf{e}_1 + \dots + y_{n+1} \mathbf{e}_{n+1}$$

is given by the superposition formula

$$\mathbf{x} = y_1 \mathbf{x}_1 + \dots + y_{n+1} \mathbf{x}_{n+1}.$$

The particular interpolation equation (5.53) corresponds to the interpolation data  $\mathbf{y} = \mathbf{e}_k$ , meaning that  $y_k = 1$ , while  $y_i = 0$  at all points  $t_i$  with  $i \neq k$ . If we can find the  $n+1$  particular interpolating polynomials that realize this very special data, we can use superposition to construct the general interpolating polynomial.

**Theorem 5.19.** Given distinct sample points  $t_1, \dots, t_{n+1}$ , the  $k^{\text{th}}$  *Lagrange interpolating polynomial* is given by

$$L_k(t) = \frac{(t - t_1) \cdots (t - t_{k-1})(t - t_{k+1}) \cdots (t - t_{n+1})}{(t_k - t_1) \cdots (t_k - t_{k-1})(t_k - t_{k+1}) \cdots (t_k - t_{n+1})}, \quad k = 1, \dots, n+1. \quad (5.54)$$

It is the unique polynomial of degree  $n$  that satisfies

$$L_k(t_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases} \quad i, k = 1, \dots, n+1. \quad (5.55)$$

*Proof:* The uniqueness of the Lagrange interpolating polynomial is an immediate consequence of Theorem 5.17. To show that (5.54) is the correct formula, we note that when  $t = t_i$  for any  $i \neq k$ , the factor  $(t - t_i)$  in the numerator of  $L_k(t)$  vanishes, while the denominator is not zero, since the points are distinct:  $t_i \neq t_k$  for  $i \neq k$ . On the other hand, when  $t = t_k$ , the numerator and denominator are equal, and so  $L_k(t_k) = 1$ . *Q.E.D.*

**Theorem 5.20.** If  $t_1, \dots, t_{n+1}$  are distinct, then the polynomial of degree  $\leq n$  that interpolates the associated data  $y_1, \dots, y_{n+1}$  is

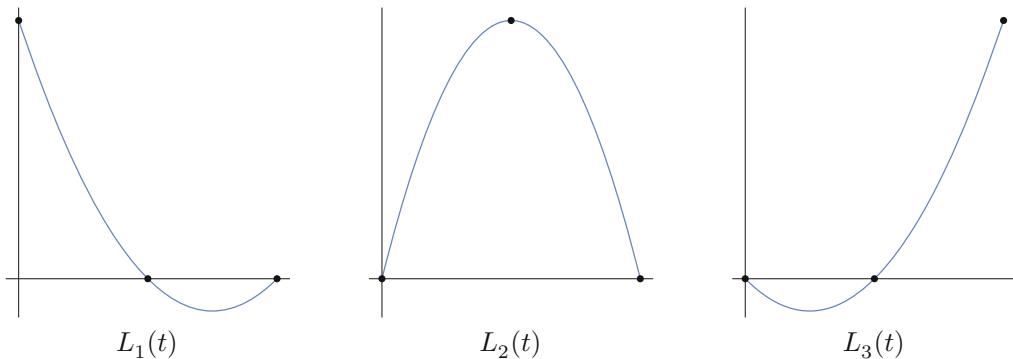
$$p(t) = y_1 L_1(t) + \dots + y_{n+1} L_{n+1}(t). \quad (5.56)$$

*Proof:* We merely compute

$$p(t_k) = y_1 L_1(t_k) + \dots + y_k L_k(t_k) + \dots + y_{n+1} L_{n+1}(t_k) = y_k,$$

where, according to (5.55), every summand except the  $k^{\text{th}}$  is zero.

*Q.E.D.*



**Figure 5.8.** Lagrange Interpolating Polynomials for the Points 0, .5, 1.

**Example 5.21.** For example, the three quadratic Lagrange interpolating polynomials for the values  $t_1 = 0$ ,  $t_2 = \frac{1}{2}$ ,  $t_3 = 1$ , used to interpolate  $e^t$  in Example 5.18 are

$$\begin{aligned} L_1(t) &= \frac{(t - \frac{1}{2})(t - 1)}{(0 - \frac{1}{2})(0 - 1)} = 2t^2 - 3t + 1, \\ L_2(t) &= \frac{(t - 0)(t - 1)}{(\frac{1}{2} - 0)(\frac{1}{2} - 1)} = -4t^2 + 4t, \\ L_3(t) &= \frac{(t - 0)(t - \frac{1}{2})}{(1 - 0)(1 - \frac{1}{2})} = 2t^2 - t. \end{aligned} \quad (5.57)$$

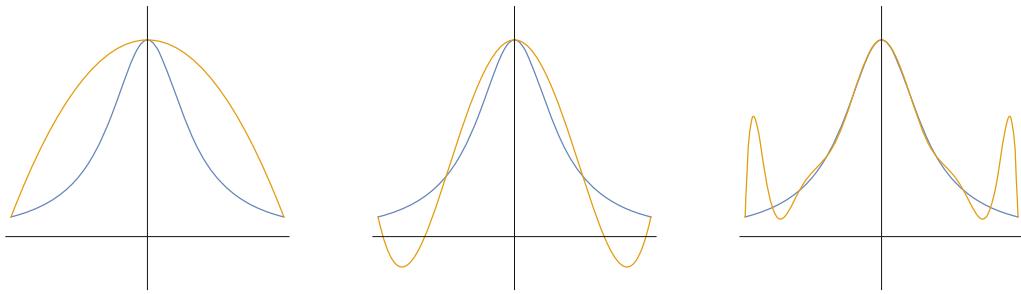
Thus, we can rewrite the quadratic interpolant (5.52) to  $e^t$  as

$$\begin{aligned} y(t) &= L_1(t) + e^{1/2} L_2(t) + e L_3(t) \\ &= (2t^2 - 3t + 1) + 1.64872(-4t^2 + 4t) + 2.71828(2t^2 - t). \end{aligned}$$

We stress that this is the *same* interpolating polynomial — we have merely rewritten it in the alternative Lagrange form.

You might expect that the higher the degree, the more accurate the interpolating polynomial. This expectation turns out, unfortunately, not to be uniformly valid. While low-degree interpolating polynomials are usually reasonable approximants to functions, not only are high-degree interpolants more expensive to compute, but they can be rather badly behaved, particularly near the ends of the interval. For example, Figure 5.9 displays the degree 2, 4, and 10 interpolating polynomials for the function  $1/(1+t^2)$  on the interval  $-3 \leq t \leq 3$  using equally spaced data points. Note the rather poor approximation of the function near the ends of the interval. Higher degree interpolants fare even worse, although the bad behavior becomes more and more concentrated near the endpoints. (Interestingly, this behavior is a consequence of the positions of the complex singularities of the function being interpolated, [62].) As a consequence, high-degree polynomial interpolation tends not to be used in practical applications. Better alternatives rely on least squares approximants by low-degree polynomials, to be described next, and interpolation by piecewise cubic splines, to be discussed at the end of this section.

If we have  $m > n + 1$  data points, then, usually, there is no degree  $n$  polynomial that fits all the data, and so we must switch over to a least squares approximation. The first



**Figure 5.9.** Degree 2, 4 and 10 Interpolating Polynomials for  $1/(1+t^2)$ .

requirement is that the associated  $m \times (n+1)$  interpolation matrix (5.50) have rank  $n+1$ ; this follows from Lemma 5.16 (coupled with Exercise 2.5.43), provided that at least  $n+1$  of the values  $t_1, \dots, t_m$  are distinct. Thus, given data at  $m \geq n+1$  distinct sample points  $t_1, \dots, t_m$ , we can uniquely determine the best least squares polynomial of degree  $n$  that best fits the data by solving the normal equations (5.42).

**Example 5.22.** Let us return to the problem of approximating the exponential function  $e^t$ . If we use more than three data points, but still require a quadratic polynomial, then we can no longer interpolate exactly, and must devise a least squares approximant. For instance, using five equally spaced sample points

$$t_1 = 0, \quad t_2 = .25, \quad t_3 = .5, \quad t_4 = .75, \quad t_5 = 1,$$

the coefficient matrix and sampled data vector (5.50) are

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .25 & .0625 \\ 1 & .5 & .25 \\ 1 & .75 & .5625 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} e^{t_1} \\ e^{t_2} \\ e^{t_3} \\ e^{t_4} \\ e^{t_5} \end{pmatrix} = \begin{pmatrix} 1. \\ 1.28403 \\ 1.64872 \\ 2.11700 \\ 2.71828 \end{pmatrix}.$$

The solution to the normal equations (5.27), with

$$K = A^T A = \begin{pmatrix} 5. & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.38281 \end{pmatrix}, \quad \mathbf{f} = A^T \mathbf{y} = \begin{pmatrix} 8.76803 \\ 5.45140 \\ 4.40153 \end{pmatrix},$$

is

$$\mathbf{x} = K^{-1} \mathbf{f} = (1.00514, .864277, .843538)^T.$$

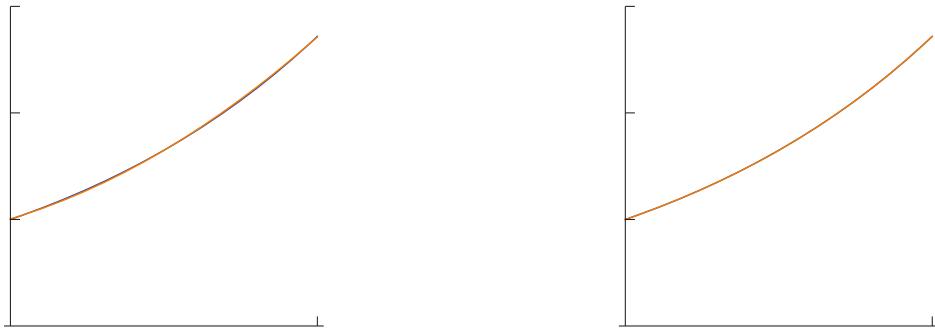
This leads to the quadratic least squares approximant

$$p_2(t) = 1.00514 + .864277t + .843538t^2.$$

On the other hand, the quartic interpolating polynomial

$$p_4(t) = 1 + .998803t + .509787t^2 + .140276t^3 + .069416t^4$$

is found directly from the data values as above. The quadratic polynomial has a maximal error of  $\approx .011$  over the interval  $[0, 1]$  — slightly better than our previous quadratic interpolant — while the quartic has a significantly smaller maximal error:  $\approx .0000527$ .



**Figure 5.10.** Quadratic Approximant and Quartic Interpolant for  $e^t$ .

(In this case, high-degree interpolants are not ill behaved.) See [Figure 5.10](#) for a comparison of the graphs, and Example 5.24 below for further discussion.

As noted above, the required calculations can be significantly simplified by the introduction of an orthogonal basis of the least squares subspace. Let us see how this works in the present situation. Given sample points  $t_1, \dots, t_m$ , let

$$\mathbf{t}_k = (t_1^k, t_2^k, \dots, t_m^k)^T, \quad k = 0, 1, 2, \dots,$$

be the vector obtained by sampling the monomial  $t^k$ . More generally, sampling a polynomial, i.e., a linear combination of monomials

$$y = p(t) = \alpha_0 + \alpha_1 t + \dots + \alpha_n t^n \quad (5.58)$$

results in the selfsame linear combination

$$\mathbf{p} = (p(t_1), \dots, p(t_n))^T = \alpha_0 \mathbf{t}_0 + \alpha_1 \mathbf{t}_1 + \dots + \alpha_n \mathbf{t}_n \quad (5.59)$$

of monomial sample vectors. Thus, all sampled polynomial vectors belong to the subspace  $W = \text{span} \{ \mathbf{t}_0, \dots, \mathbf{t}_n \} \subset \mathbb{R}^m$  spanned by the monomial sample vectors.

Let  $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$  be data that has been measured at the sample points. The polynomial least squares approximation is, by definition, the polynomial  $y = p(t)$  whose sample vector  $\mathbf{p}$  is the closest point to  $\mathbf{y}$  lying in the subspace  $W$ , which, according to Theorem 5.7, is the same as the orthogonal projection of the data vector  $\mathbf{y}$  onto  $W$ . But the monomial sample vectors  $\mathbf{t}_0, \dots, \mathbf{t}_n$  are not orthogonal, and so a direct approach requires solving the normal equations (5.42) for the least squares coefficients  $\alpha_0, \dots, \alpha_n$ .

A better strategy is to first apply the Gram–Schmidt process to construct an orthogonal basis for the subspace  $W$ , from which the least squares coefficients are then found by our orthogonal projection formula (4.41). Let us adopt the rescaled version

$$\langle \mathbf{v}, \mathbf{w} \rangle = \frac{1}{m} \sum_{i=1}^m v_i w_i = \overline{v w} \quad (5.60)$$

of the standard dot product<sup>†</sup> on  $\mathbb{R}^m$ . If  $\mathbf{v}, \mathbf{w}$  represent the sample vectors corresponding to the functions  $v(t), w(t)$ , then their inner product  $\langle \mathbf{v}, \mathbf{w} \rangle$  is equal to the average value

---

<sup>†</sup> For weighted least squares, we would use an appropriately weighted inner product.

of the product function  $v(t)w(t)$  on the  $m$  sample points. In particular, the inner product between our “monomial” basis vectors corresponding to sampling  $t^k$  and  $t^l$  is

$$\langle \mathbf{t}_k, \mathbf{t}_l \rangle = \frac{1}{m} \sum_{i=1}^m t_i^k t_i^l = \frac{1}{m} \sum_{i=1}^m t_i^{k+l} = \overline{t^{k+l}}, \quad (5.61)$$

which is the averaged sample value of the monomial  $t^{k+l}$ .

To keep the formulas reasonably simple, let us further assume<sup>†</sup> that the sample points are evenly spaced and symmetric about 0. The first requirement is that  $t_i - t_{i-1} = h$  be independent of  $i$ , while the second means that if  $t_i$  is a sample point, then so is  $-t_i$ . An example is provided by the seven sample points  $-3, -2, -1, 0, 1, 2, 3$ . As a consequence of these two assumptions, the averaged sample values of the odd powers of  $t$  vanish:  $\overline{t^{2i+1}} = 0$ . Hence, by (5.61), the sample vectors  $\mathbf{t}_k$  and  $\mathbf{t}_l$  are orthogonal whenever  $k + l$  is odd.

Applying the Gram–Schmidt algorithm to  $\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \dots$  produces the orthogonal basis vectors  $\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2, \dots$ . Each

$$\mathbf{q}_k = (q_k(t_1), \dots, q_k(t_m))^T = c_{k0} \mathbf{t}_0 + c_{k1} \mathbf{t}_1 + \dots + c_{kk} \mathbf{t}_k \quad (5.62)$$

can be interpreted as the sample vector for a certain degree  $k$  interpolating polynomial

$$q_k(t) = c_{k0} + c_{k1} t + \dots + c_{kk} t^k.$$

Under these assumptions, the first few of these polynomials, along with their corresponding orthogonal sample vectors, are as follows:

$$\begin{aligned} q_0(t) &= 1, & \mathbf{q}_0 &= \mathbf{t}_0, & \|\mathbf{q}_0\|^2 &= 1, \\ q_1(t) &= t, & \mathbf{q}_1 &= \mathbf{t}_1, & \|\mathbf{q}_1\|^2 &= \overline{t^2}, \\ q_2(t) &= t^2 - \overline{t^2}, & \mathbf{q}_2 &= \mathbf{t}_2 - \overline{t^2} \mathbf{t}_0, & \|\mathbf{q}_2\|^2 &= \overline{t^4} - (\overline{t^2})^2, \\ q_3(t) &= t^3 - \frac{\overline{t^4}}{\overline{t^2}} t, & \mathbf{q}_3 &= \mathbf{t}_3 - \frac{\overline{t^4}}{\overline{t^2}} \mathbf{t}_1, & \|\mathbf{q}_3\|^2 &= \overline{t^6} - \frac{(\overline{t^4})^2}{\overline{t^2}}. \end{aligned} \quad (5.63)$$

With these in hand, the least squares approximating polynomial of degree  $n$  to the given data vector  $\mathbf{y}$  is given by a linear combination

$$p(t) = a_0 q_0(t) + a_1 q_1(t) + a_2 q_2(t) + \dots + a_n q_n(t). \quad (5.64)$$

The coefficients can now be obtained directly through the orthogonality formula (4.42), and so

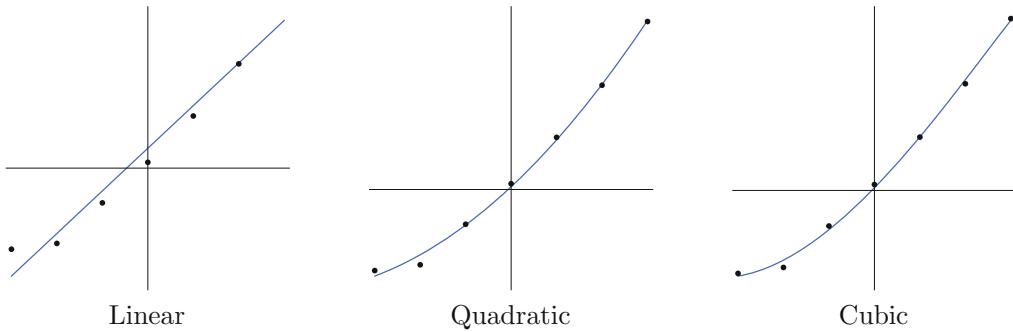
$$a_k = \frac{\langle \mathbf{q}_k, \mathbf{y} \rangle}{\|\mathbf{q}_k\|^2} = \frac{\overline{q_k y}}{\overline{q_k^2}}. \quad (5.65)$$

Thus, once we have set up the orthogonal basis, we no longer need to solve any linear system to construct the least squares approximation.

An additional advantage of the orthogonal basis is that, in contrast to the direct method, the formulas (5.65) for the least squares coefficients *do not depend on the degree of the approximating polynomial*. As a result, one can readily increase the degree, and, in favorable

---

<sup>†</sup> The method works without this restriction, but the formulas become more unwieldy. See Exercise 5.5.31 for details.



**Figure 5.11.** Least Squares Data Approximations.

situations, the accuracy of the approximant without having to recompute any of the lower-degree terms. For instance, if a quadratic polynomial  $p_2(t) = a_0 + a_1 q_1(t) + a_2 q_2(t)$  is insufficiently accurate, the cubic least squares approximant  $p_3(t) = p_2(t) + a_3 q_3(t)$  can be constructed *without* having to recompute the quadratic coefficients  $a_0, a_1, a_2$ . This doesn't work when using the non-orthogonal monomials, all of whose coefficients will be affected by increasing the degree of the approximating polynomial.

**Example 5.23.** Consider the following tabulated sample values:

$t_i$	-3	-2	-1	0	1	2	3
$y_i$	-1.4	-1.3	-0.6	0.1	0.9	1.8	2.9

To compute polynomial least squares fits of degrees 1, 2 and 3, we begin by computing the polynomials (5.63), which for the given sample points  $t_i$  are

$$\begin{aligned} q_0(t) &= 1, & q_1(t) &= t, & q_2(t) &= t^2 - 4, & q_3(t) &= t^3 - 7t, \\ \|\mathbf{q}_0\|^2 &= 1, & \|\mathbf{q}_1\|^2 &= 4, & \|\mathbf{q}_2\|^2 &= 12, & \|\mathbf{q}_3\|^2 &= \frac{216}{7}. \end{aligned}$$

Thus, to four decimal places, the coefficients (5.64) for the least squares approximation are

$$\begin{aligned} a_0 &= \langle \mathbf{q}_0, \mathbf{y} \rangle = .3429, & a_1 &= \frac{1}{4} \langle \mathbf{q}_1, \mathbf{y} \rangle = .7357, \\ a_2 &= \frac{1}{12} \langle \mathbf{q}_2, \mathbf{y} \rangle = .0738, & a_3 &= \frac{7}{216} \langle \mathbf{q}_3, \mathbf{y} \rangle = -.0083. \end{aligned}$$

To obtain the best linear approximation, we use

$$p_1(t) = a_0 q_0(t) + a_1 q_1(t) = .3429 + .7357t,$$

with a least squares error of .7081. Similarly, the quadratic and cubic least squares approximations are

$$\begin{aligned} p_2(t) &= .3429 + .7357t + .0738(t^2 - 4), \\ p_3(t) &= .3429 + .7357t + .0738(t^2 - 4) - .0083(t^3 - 7t), \end{aligned}$$

with respective least squares errors .2093 and .1697 at the sample points. Observe that, as noted above, the lower order coefficients do not change as we increase the degree of the approximating polynomial. A plot of the first three approximations appears in [Figure 5.11](#).

The small cubic term does not significantly increase the accuracy of the approximation, and so this data probably comes from sampling a quadratic function.

*Proof of Lemma 5.16:* We will establish the rather striking *LU* factorization of the transposed Vandermonde matrix  $V = A^T$ , which will immediately prove that, when  $t_1, \dots, t_{n+1}$  are distinct, both  $V$  and  $A$  are nonsingular matrices. The  $4 \times 4$  case is instructive for understanding the general pattern. Applying regular Gaussian Elimination, we obtain the explicit *LU* factorization

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ t_1 & t_2 & t_3 & t_4 \\ t_1^2 & t_2^2 & t_3^2 & t_4^2 \\ t_1^3 & t_2^3 & t_3^3 & t_4^3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ t_1 & 1 & 0 & 0 \\ t_1^2 & t_1 + t_2 & 1 & 0 \\ t_1^3 & t_1^2 + t_1 t_2 + t_2^2 & t_1 + t_2 + t_3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & t_2 - t_1 & t_3 - t_1 & t_4 - t_1 \\ 0 & 0 & (t_3 - t_1)(t_3 - t_2) & (t_4 - t_1)(t_4 - t_2) \\ 0 & 0 & 0 & (t_4 - t_1)(t_4 - t_2)(t_4 - t_3) \end{pmatrix}.$$

Observe that the pivots, i.e., the diagonal entries of  $U$ , are all nonzero if the points  $t_1, t_2, t_3, t_4$  are distinct. The reader may be able to spot the pattern in the above formula and thus guess the general case. Indeed, the individual entries of the matrices appearing in the factorization

$$V = LU \tag{5.66}$$

of the  $(n+1) \times (n+1)$  Vandermonde matrix are

$$\begin{aligned} v_{ij} &= t_j^{i-1}, & i, j = 1, \dots, n+1, \\ \ell_{ij} &= \sum_{1 \leq k_1 \leq \dots \leq k_{i-j} \leq j} t_{k_1} t_{k_2} \cdots t_{k_{i-j}}, & 1 \leq j < i \leq n+1, \\ u_{ij} &= \prod_{k=1}^i (t_j - t_k), & 1 < i \leq j \leq n+1, \\ & & \ell_{ii} = 1, \quad i = 1, \dots, n+1, \\ & & \ell_{ij} = 0, \quad 1 \leq i < j \leq n+1, \\ & & u_{1j} = 1, \quad j = 1, \dots, n+1, \\ & & u_{ij} = 0, \quad 1 \leq j < i \leq n+1. \end{aligned} \tag{5.67}$$

Full details of the proof that (5.67) holds can be found in [30, 63]. (Surprisingly, as far as we know, these are the first places this factorization appears in the literature.) The entries of  $L$  lying below the diagonal are known as the *complete monomial polynomials* since  $\ell_{ij}$  is obtained by summing, with unit coefficients, all monomials of degree  $i-j$  in the  $j$  variables  $t_1, \dots, t_j$ . The entries of  $U$  appearing on or above the diagonal are known as the *Newton difference polynomials*. In particular, if  $t_1, \dots, t_n$  are distinct, so  $t_i \neq t_j$  for  $i \neq j$ , then all entries of  $U$  lying on or above the diagonal are nonzero. In this case,  $V$  has all nonzero pivots, namely, the diagonal entries of  $U$ , and is a regular, hence nonsingular matrix.

*Q.E.D.*

## Exercises

- 5.5.13. Find and graph the polynomial of minimal degree that passes through the following points: (a)  $(3, -1), (6, 5)$ ; (b)  $(-2, 4), (0, 6), (1, 10)$ ; (c)  $(-2, 3), (0, -1), (1, -3)$ ; (d)  $(-1, 2), (0, -1), (1, 0), (2, -1)$ ; (e)  $(-2, 17), (-1, -3), (0, -3), (1, -1), (2, 9)$ .

5.5.14. For the following data values, construct the interpolating polynomial in Lagrange form:

$t_i$	-3	2
$y_i$	1	5

$t_i$	0	1	3
$y_i$	1	.5	.25

$t_i$	-1	0	1
$y_i$	1	2	-1

$t_i$	0	1	2	3
$y_i$	0	1	4	9

$t_i$	-2	-1	0	1	2
$y_i$	-1	-2	2	1	3

5.5.15. Given  $\begin{array}{c|ccc} t_i & 1 & 2 & 3 \\ \hline y_i & 3 & 6 & 11 \end{array}$  (a) find the straight line  $y = \alpha + \beta t$  that best fits the data in the least squares sense; (b) find the parabola  $y = \alpha + \beta t + \gamma t^2$  that best fits the data. Interpret the error.

5.5.16. Re-solve Exercise 5.5.15 using the respective weights 2, 1, .5 at the three data points.

5.5.17. The table  $\begin{array}{c|cccc} \text{time in sec} & 0 & 10 & 20 & 30 \\ \hline \text{meters} & 4500 & 4300 & 3930 & 3000 \end{array}$  measures the altitude

of a falling parachutist before her chute has opened. Predict how many seconds she can wait before reaching the minimum altitude of 1500 meters.

5.5.18. A missile is launched in your direction. Using a range finder, you measure its altitude

at the times:	time in sec	0	10	20	30	40	50
	altitude in meters	200	650	970	1200	1375	1130

How long until you have to run?

5.5.19. A student runs an experiment six times in an attempt to obtain an equation relating two physical quantities  $x$  and  $y$ . For  $x = 1, 2, 4, 6, 8, 10$  units, the experiments result in corresponding  $y$  values of 3, 3, 4, 6, 7, 8 units. Find and graph the following: (a) the least squares line; (b) the least squares quadratic polynomial; (c) the interpolating polynomial. (d) Which do you think is the most likely theoretical model for this data?

5.5.20. (a) Write down the Taylor polynomials of degrees 2 and 4 at  $t = 0$  for the function  $f(t) = e^t$ . (b) Compare their accuracy with the interpolating and least squares polynomials in Examples 5.18 and 5.22.

◇ 5.5.21. Given the values of  $\sin t$  at  $t = 0^\circ, 30^\circ, 45^\circ, 60^\circ$ , find the following approximations: (a) the least squares linear polynomial; (b) the least squares quadratic polynomial; (c) the quadratic Taylor polynomial at  $t = 0$ ; (d) the interpolating polynomial; (e) the cubic Taylor polynomial at  $t = 0$ ; (f) Graph each approximation and discuss its accuracy.

5.5.22. Find the quartic (degree 4) polynomial that exactly interpolates the function  $\tan t$  at the five data points  $t_0 = 0, t_1 = .25, t_2 = .5, t_3 = .75, t_4 = 1$ . Compare the graphs of the two functions over  $0 \leq t \leq \frac{1}{2}\pi$ .

5.5.23. (a) Find the least squares linear polynomial approximating  $\sqrt{t}$  on  $[0, 1]$ , choosing six different exact data values. (b) How much more accurate is the least squares quadratic polynomial based on the same data?

5.5.24. A table of logarithms contains the following entries:

$t$	1.0	2.0	3.0	4.0
$\log_{10} t$	0	.3010	.4771	.6021

Approximate  $\log_{10} e$  by constructing an interpolating polynomial of (a) degree two using the entries at  $x = 1.0, 2.0$ , and  $3.0$ , (b) degree three using all the entries.

- ◇ 5.5.25. Let  $q(t)$  denote the quadratic interpolating polynomial that goes through the data points  $(t_0, y_0), (t_1, y_1), (t_2, y_2)$ . (a) Under what conditions does  $q(t)$  have a minimum? A maximum? (b) Show that the minimizing/maximizing value is at  $t^* = \frac{m_0 s_1 - m_1 s_0}{s_1 - s_0}$ , where  $s_0 = \frac{y_1 - y_0}{t_1 - t_0}$ ,  $s_1 = \frac{y_2 - y_1}{t_2 - t_1}$ ,  $m_0 = \frac{t_0 + t_1}{2}$ ,  $m_1 = \frac{t_1 + t_2}{2}$ . (c) What is  $q(t^*)$ ?

- 5.5.26. Use the orthogonal sample vectors (5.63) to find the best polynomial least squares fits of degree 1, 2 and 3 for the following sets of data:

(a)	$t_i$	-2	-1	0	1	2		
	$y_i$	7	11	13	18	21		
(b)	$t_i$	-3	-2	-1	0	1	2	3
	$y_i$	-2.7	-2.1	-.5	.5	1.2	2.4	3.2
(c)	$t_i$	-3	-2	-1	0	1	2	3
	$y_i$	60	80	90	100	120	120	130

- ◇ 5.5.27. (a) Verify the orthogonality of the sample polynomial vectors in (5.63). (b) Construct the next orthogonal sample polynomial  $q_4(t)$  and the norm of its sample vector. (c) Use your result to compute the quartic least squares approximation for the data in Example 5.23.

- 5.5.28. Use the result of Exercise 5.5.27 to find the best approximating polynomial of degree 4 to the data in Exercise 5.5.26.

- 5.5.29. Justify the fact that the orthogonal sample vector  $\mathbf{q}_k$  in (5.62) is a linear combination of only the first  $k$  monomial sample vectors.

- ◇ 5.5.30. The formulas (5.63) apply only when the sample times are symmetric around 0. When the sample points  $t_1, \dots, t_n$  are equally spaced, so  $t_{i+1} - t_i = h$  for all  $i = 1, \dots, n-1$ , then there is a simple trick to convert the least squares problem into a symmetric form.

- (a) Show that the translated sample points  $s_i = t_i - \bar{t}$ , where  $\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$  is the average, are symmetric around 0. (b) Suppose  $q(s)$  is the least squares polynomial for the data points  $(s_i, y_i)$ . Prove that  $p(t) = q(t - \bar{t})$  is the least squares polynomial for the original data  $(t_i, y_i)$ . (c) Apply this method to find the least squares polynomials of degrees 1 and

2 for the following data:	$t_i$	1	2	3	4	5	6
	$y_i$	-8	-6	-4	-1	1	3

- ◇ 5.5.31. Construct the first three orthogonal basis elements for sample points  $t_1, \dots, t_m$  that are in general position.

- ♣ 5.5.32. Use  $n+1$  equally spaced data points to interpolate  $f(t) = 1/(1+t^2)$  on an interval  $-a \leq t \leq a$  for  $a = 1, 1.5, 2, 2.5, 3$  and  $n = 2, 4, 10, 20$ . Do all intervals exhibit the pathology illustrated in Figure 5.9? If not, how large can  $a$  be before the interpolants have poor approximation properties? What happens when the number of interpolation points is taken to be  $n = 50$ ?

- ♣ 5.5.33. Repeat Exercise 5.5.32 for the hyperbolic secant function  $f(t) = \operatorname{sech} t = 1/\cosh t$ .

- 5.5.34. Given  $A$  as in (5.50) with  $m < n+1$ , how would you characterize those polynomials  $p(t)$  whose coefficient vector  $\mathbf{x}$  lies in  $\ker A$ ?

- 5.5.35. (a) Give an example of an interpolating polynomial through  $n+1$  points that has degree  $< n$ . (b) Can you explain, without referring to the explicit formulas, why all the Lagrange interpolating polynomials based on  $n+1$  points must have degree equal to  $n$ ?

5.5.36. Let  $x_1, \dots, x_n$  be distinct real numbers. Prove that the  $n \times n$  matrix  $K$  with entries

$$k_{ij} = \frac{1 - (x_i x_j)^n}{1 - x_i x_j}$$

is positive definite.

◇ 5.5.37. Prove the determinant formula  $\det A = \prod_{1 \leq i < j \leq n+1} (t_i - t_j)$  for the  $(n+1) \times (n+1)$  Vandermonde matrix defined in (5.50).

◇ 5.5.38. (a) Prove that a polynomial  $p(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$  of degree  $\leq n$  vanishes at  $n+1$  *distinct* points, so  $p(x_1) = p(x_2) = \dots = p(x_{n+1}) = 0$ , if and only if  $p(x) \equiv 0$  is the zero polynomial. (b) Prove that the monomials  $1, x, x^2, \dots, x^n$  are linearly independent. (c) Explain why  $p(x) \equiv 0$  if and only if all its coefficients  $a_0 = a_1 = \dots = a_n = 0$ .

*Hint:* Use Lemma 5.16 and Exercise 2.3.37.

◇ 5.5.39. *Numerical differentiation:* The most common numerical methods for approximating the derivatives of a function are based on interpolation. To approximate the  $k^{\text{th}}$  derivative  $f^{(k)}(x_0)$  at a point  $x_0$ , one replaces the function  $f(x)$  by an interpolating polynomial  $p_n(x)$  of degree  $n \geq k$  based on the nearby points  $x_0, \dots, x_n$  (the point  $x_0$  is almost always included as an interpolation point), leading to the approximation  $f^{(k)}(x_0) \approx p_n^{(k)}(x_0)$ . Use this method to construct numerical approximations to (a)  $f'(x)$  using a quadratic interpolating polynomial based on  $x - h, x, x + h$ . (b)  $f''(x)$  with the same quadratic polynomial. (c)  $f'(x)$  using a quadratic interpolating polynomial based on  $x, x + h, x + 2h$ . (d)  $f'(x), f''(x), f'''(x)$  and  $f^{(iv)}(x)$  using a quartic interpolating polynomial based on  $x - 2h, x - h, x, x + h, x + 2h$ . (e) Test your methods by approximating the derivatives of  $e^x$  and  $\tan x$  at  $x = 0$  with step sizes  $h = \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}, \frac{1}{10000}$ . Discuss the accuracies you observe. Can the step size be arbitrarily small? (f) Why do you need  $n \geq k$ ?

◇ 5.5.40. *Numerical integration:* Most numerical methods for evaluating a definite integral  $\int_a^b f(x) dx$  are based on interpolation. One chooses  $n+1$  interpolation points  $a \leq x_0 < x_1 < \dots < x_n \leq b$  and replaces the integrand by its interpolating polynomial  $p_n(x)$  of degree  $n$ , leading to the approximation  $\int_a^b f(x) dx \approx \int_a^b p_n(x) dx$ , where the polynomial integral can be done explicitly. Write down the following popular integration rules:

- (a) *Trapezoid Rule:*  $x_0 = a, x_1 = b$ . (b) *Simpson's Rule:*  $x_0 = a, x_1 = \frac{1}{2}(a+b), x_2 = b$ .
- (c) *Simpson's  $\frac{3}{8}$  Rule:*  $x_0 = a, x_1 = \frac{1}{3}(a+b), x_2 = \frac{2}{3}(a+b), x_3 = b$ .
- (d) *Midpoint Rule:*  $x_0 = \frac{1}{2}(a+b)$ . (e) *Open Rule:*  $x_0 = \frac{1}{3}(a+b), x_1 = \frac{2}{3}(a+b)$ .
- (f) Test your methods for accuracy on the following integrals:

$$(i) \int_0^1 e^x dx, \quad (ii) \int_0^\pi \sin x dx, \quad (iii) \int_1^e \log x dx, \quad (iv) \int_0^{\pi/2} \sqrt{x^3 + 1} dx.$$

*Note:* For more details on numerical differentiation and integration, you are encouraged to consult a basic numerical analysis text, e.g., [8].

## Approximation and Interpolation by General Functions

There is nothing special about polynomial functions in the preceding approximation and interpolation schemes. For example, suppose we are interested in determining the best trigonometric approximation

$$y = \alpha_1 \cos t + \alpha_2 \sin t$$

to a given set of data. Again, the least squares error takes the same form as in (5.49),

namely  $\|\mathbf{e}\|^2 = \|\mathbf{y} - A\mathbf{x}\|^2$ , where

$$A = \begin{pmatrix} \cos t_1 & \sin t_1 \\ \cos t_2 & \sin t_2 \\ \vdots & \vdots \\ \cos t_m & \sin t_m \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Thus, the columns of  $A$  are the sampled values of the functions  $\cos t, \sin t$ . The key requirement is that the unspecified parameters — in this case  $\alpha_1, \alpha_2$  — occur *linearly* in the approximating function. Thus, the most general case is to approximate the data (5.38) by a linear combination

$$y(t) = \alpha_1 h_1(t) + \alpha_2 h_2(t) + \cdots + \alpha_n h_n(t)$$

of prescribed functions  $h_1(x), \dots, h_n(x)$ . The total squared error is, as always, given by

$$\text{Squared Error} = \sum_{i=1}^m [y_i - y(t_i)]^2 = \|\mathbf{y} - A\mathbf{x}\|^2,$$

where the sample matrix  $A$ , the vector of unknown coefficients  $\mathbf{x}$ , and the data vector  $\mathbf{y}$  are

$$A = \begin{pmatrix} h_1(t_1) & h_2(t_1) & \dots & h_n(t_1) \\ h_1(t_2) & h_2(t_2) & \dots & h_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(t_m) & h_2(t_m) & \dots & h_n(t_m) \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}. \quad (5.68)$$

If  $A$  is square and nonsingular, then we can find an interpolating function of the prescribed form by solving the linear system

$$A\mathbf{x} = \mathbf{y}. \quad (5.69)$$

A particularly important case is provided by the  $2n + 1$  trigonometric functions

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x, \quad \dots \quad \cos nx, \quad \sin nx.$$

Interpolation on  $2n + 1$  equally spaced data points on the interval  $[0, 2\pi]$  leads to the Discrete Fourier Transform, used in signal processing, data transmission, and compression, and to be the focus of Section 5.6.

If there are more than  $n$  data points, then we cannot, in general, interpolate exactly, and must content ourselves with a least squares approximation that minimizes the error at the sample points as best it can. The least squares solution to the interpolation equations (5.69) is found by solving the associated normal equations  $K\mathbf{x} = \mathbf{f}$ , where the  $(i, j)$  entry of  $K = A^T A$  is  $m$  times the average sample value of the product of  $h_i(t)$  and  $h_j(t)$ , namely

$$k_{ij} = m \overline{h_i(t) h_j(t)} = \sum_{\ell=1}^m h_i(t_\ell) h_j(t_\ell), \quad (5.70)$$

whereas the  $i^{\text{th}}$  entry of  $\mathbf{f} = A^T \mathbf{y}$  is

$$f_i = m \overline{h_i(t) y} = \sum_{\ell=1}^m h_i(t_\ell) y_\ell. \quad (5.71)$$

The one issue is whether the columns of the sample matrix  $A$  are linearly independent, which is a more subtle issue than the polynomial case covered by Lemma 5.16. Linear independence of the sampled function vectors is, in general, more restrictive than merely requiring the functions themselves to be linearly independent; see Exercise 2.3.37 for details.

If the parameters do not occur linearly in the functional formula, then we cannot use linear algebra to effect a least squares approximation. For example, one cannot use least squares to determine the frequency  $\omega$ , the amplitude  $r$ , and the phase shift  $\delta$  of the general trigonometric approximation

$$y = c_1 \cos \omega t + c_2 \sin \omega t = r \cos(\omega t + \delta)$$

that minimizes the least squares error at the sample points. Approximating data by such a function constitutes a *nonlinear* optimization problem.

## Exercises

5.5.41. Given the values 

$t_i$	0	.5	1
$y_i$	1	.5	.25

 construct the trigonometric function of

the form  $g(t) = a \cos \pi t + b \sin \pi t$  that best approximates the data in the least squares sense.

5.5.42. Find the hyperbolic function  $g(t) = a \cosh t + b \sinh t$  that best approximates the data in Exercise 5.5.41.

5.5.43. (a) Find the exponential function of the form  $g(t) = a e^t + b e^{2t}$  that best approximates  $t^2$  in the least squares sense based on the sample points 0, 1, 2, 3, 4. (b) What is the least squares error? (c) Compare the graphs on the interval  $[0, 4]$  — where is the approximation the worst? (d) How much better can you do by including a constant term in  $g(t) = a e^t + b e^{2t} + c$ ?

5.5.44. (a) Find the best trigonometric approximation of the form  $g(t) = r \cos(t + \delta)$  to  $t^2$  using 5 and 9 equally spaced sample points on  $[0, \pi]$ .

(b) Can you answer the question for  $g(t) = r_1 \cos(t + \delta_1) + r_2 \cos(2t + \delta_2)$ ?

♡ 5.5.45. A *trigonometric polynomial* of degree  $n$  is a function of the form

$$p(t) = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t + \cdots + a_n \cos nt + b_n \sin nt,$$

where  $a_0, a_1, b_1, \dots, a_n, b_n$  are the coefficients. Find the trigonometric polynomial of degree  $n$  that is the least squares approximation to the function  $f(t) = 1/(1 + t^2)$  on the interval

$$[-\pi, \pi] \text{ based on the } k \text{ equally spaced data points } t_j = -\pi + \frac{2\pi j}{k}, \quad j = 0, \dots, k-1,$$

(omitting the right-hand endpoint), when (a)  $n = 1$ ,  $k = 4$ , (b)  $n = 2$ ,  $k = 8$ , (c)  $n = 2$ ,  $k = 16$ , (d)  $n = 3$ ,  $k = 16$ . Compare the graphs of the trigonometric approximant and the function, and discuss. (e) Why do we not include the right-hand endpoint  $t_k = \pi$ ?

♡ 5.5.46. The *sinc functions* are defined as  $S_0(x) = \frac{\sin(\pi x/h)}{\pi x/h}$ , while  $S_j(x) = S_0(x - jh)$

whenever  $h > 0$  and  $j$  is an integer. We will interpolate a function  $f(x)$  at the mesh points  $x_j = jh$ ,  $j = 0, \dots, n$ , by a linear combination of sinc functions:  $S(x) = c_0 S_0(x) + \cdots + c_n S_n(x)$ . What are the coefficients  $c_j$ ? Graph and discuss the accuracy of the sinc interpolant for the functions  $x^2$  and  $\frac{1}{2} - |x - \frac{1}{2}|$  on the interval  $[0, 1]$  using  $h = .25, .1$ , and  $.025$ .

## Least Squares Approximation in Function Spaces

So far, while we have used least squares minimization to interpolate and approximate known, complicated functions by simpler polynomials, we have only worried about the errors committed at a discrete, preassigned set of sample points. A more uniform approach would be to take into account the errors at *all* points in the interval of interest. This can be accomplished by replacing the discrete, finite-dimensional vector space norm on sample vectors by a continuous, infinite-dimensional function space norm in order to specify the least squares error that must be minimized over the entire interval.

More precisely, we let  $V = C^0[a, b]$  denote the space of continuous functions on the bounded interval  $[a, b]$  with  $L^2$  inner product and norm

$$\langle f, g \rangle = \int_a^b f(t) g(t) dt, \quad \|f\| = \sqrt{\int_a^b f(t)^2 dt}. \quad (5.72)$$

Let  $\mathcal{P}^{(n)} \subset C^0[a, b]$  denote the subspace consisting of all polynomials of degree  $\leq n$ . For simplicity, we employ the standard monomial basis  $1, t, t^2, \dots, t^n$ . We will be approximating a general function  $f(t) \in C^0[a, b]$  by a polynomial

$$p(t) = \alpha_0 + \alpha_1 t + \dots + \alpha_n t^n \in \mathcal{P}^{(n)} \quad (5.73)$$

of degree at most  $n$ . The *error function*  $e(t) = f(t) - p(t)$  measures the discrepancy between the function and its approximating polynomial at each  $t$ . Instead of summing the squares of the errors at a finite set of sample points, we go to a continuous limit that sums or, rather, integrates the squared errors of all points in the interval. Thus, the approximating polynomial will be characterized as the one that minimizes the total  $L^2$  *squared error*:

$$\text{Squared Error} = \|e\|^2 = \|p - f\|^2 = \int_a^b [p(t) - f(t)]^2 dt. \quad (5.74)$$

To solve the problem of minimizing the squared error, we begin by substituting (5.73) into (5.74) and expanding, as in (5.20):

$$\begin{aligned} \|p - f\|^2 &= \left\| \sum_{i=0}^n \alpha_i t^i - f(t) \right\|^2 = \left\langle \sum_{i=0}^n \alpha_i t^i - f(t), \sum_{i=0}^n \alpha_i t^i - f(t) \right\rangle \\ &= \sum_{i,j=0}^n \alpha_i \alpha_j \langle t^i, t^j \rangle - 2 \sum_{i=0}^n \alpha_i \langle t^i, f(t) \rangle + \|f(t)\|^2. \end{aligned}$$

As a result, we are required to minimize a quadratic function of the standard form

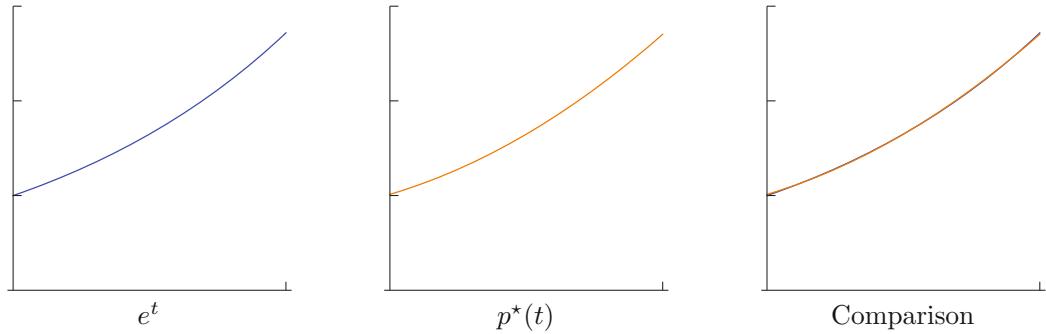
$$\mathbf{x}^T K \mathbf{x} - 2 \mathbf{x}^T \mathbf{f} + c, \quad (5.75)$$

where  $\mathbf{x} = (\alpha_0, \alpha_1, \dots, \alpha_n)^T$  is the vector containing the unknown coefficients in the minimizing polynomial (5.73), while<sup>†</sup>

$$k_{ij} = \langle t^i, t^j \rangle = \int_a^b t^{i+j} dt, \quad f_i = \langle t^i, f \rangle = \int_a^b t^i f(t) dt, \quad (5.76)$$

---

<sup>†</sup> Here, the indices  $i, j$  labeling the entries of the  $(n+1) \times (n+1)$  matrix  $K$  and vectors  $\mathbf{x}, \mathbf{f} \in \mathbb{R}^{n+1}$  range from 0 to  $n$  instead of 1 to  $n+1$ .



**Figure 5.12.** Quadratic Least Squares Approximation of  $e^t$ .

are the Gram matrix  $K$  consisting of inner products between basis monomials along with the vector  $\mathbf{f}$  of inner products between the monomials and the given function. The coefficients of the least squares minimizing polynomial are thus found by solving the associated normal equations  $K\mathbf{x} = \mathbf{f}$ .

**Example 5.24.** Let us return to the problem of approximating the exponential function  $f(t) = e^t$  by a quadratic polynomial on the interval  $0 \leq t \leq 1$ , but now with the least squares error being measured by the  $L^2$  norm. Thus, we consider the subspace  $\mathcal{P}^{(2)}$  consisting of all quadratic polynomials

$$p(t) = \alpha + \beta t + \gamma t^2.$$

Using the monomial basis  $1, t, t^2$ , the coefficient matrix is the Gram matrix  $K$  consisting of the inner products

$$\langle t^i, t^j \rangle = \int_0^1 t^{i+j} dt = \frac{1}{i+j+1}$$

between basis monomials, while the right-hand side is the vector of inner products

$$\langle t^i, e^t \rangle = \int_0^1 t^i e^t dt.$$

The solution to the normal system

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} e - 1 \\ 1 \\ e - 2 \end{pmatrix}.$$

is computed to be

$$\alpha = 39e - 105 \simeq 1.012991, \quad \beta = -216e + 588 \simeq .851125, \quad \gamma = 210e - 570 \simeq .839184,$$

leading to the least squares quadratic approximant

$$p^*(t) = 1.012991 + .851125 t + .839184 t^2, \quad (5.77)$$

plotted in [Figure 5.12](#). The least squares error is

$$\|e^t - p^*(t)\|^2 \simeq .000027835.$$

The maximal error over the interval is measured by the  $L^\infty$  norm of the difference:

$$\|e^t - p^*(t)\|_\infty = \max \{ |e^t - p^*(t)| \mid 0 \leq t \leq 1 \} \simeq .014981815,$$

with the maximum occurring at  $t = 1$ . Thus, the simple quadratic polynomial (5.77) will give a reasonable approximation to the first two decimal places in  $e^t$  on the entire interval  $[0, 1]$ . A more accurate approximation can be constructed by taking a higher degree polynomial, or by decreasing the length of the interval.

**Remark.** Although the least squares polynomial (5.77) minimizes the  $L^2$  norm of the error, it does slightly worse with the  $L^\infty$  norm than the previous sample-based minimizer (5.52). The problem of finding the quadratic polynomial that minimizes the  $L^\infty$  norm is more difficult, and must be solved by nonlinear minimization techniques.

**Remark.** As noted in Example 3.39, the Gram matrix for the simple monomial basis is the  $n \times n$  Hilbert matrix (1.72). The ill-conditioned nature of the Hilbert matrix and the consequential difficulty in accurately solving the normal equations complicate the practical numerical implementation of high-degree least squares polynomial approximations. A better approach, based on an orthogonal polynomial basis, will be discussed next.

## Exercises

- 5.5.47. Approximate the function  $f(t) = \sqrt[3]{t}$  using the least squares method based on the  $L^2$  norm on the interval  $[0, 1]$  by (a) a straight line; (b) a parabola; (c) a cubic polynomial.
- 5.5.48. Approximate the function  $f(t) = \frac{1}{8}(2t - 1)^3 + \frac{1}{4}$  by a quadratic polynomial on the interval  $[-1, 1]$  using the least squares method based on the  $L^2$  norm. Compare the graphs. Where is the error the largest?
- 5.5.49. For the function  $f(t) = \sin t$  determine the approximating linear and quadratic polynomials that minimize the least squares error based on the  $L^2$  norm on  $[0, \frac{1}{2}\pi]$ .
- 5.5.50. Find the quartic (degree 4) polynomial that best approximates the function  $e^t$  on the interval  $[0, 1]$  by minimizing the  $L^2$  error (5.72).
- ◇ 5.5.51. (a) Find the quadratic interpolant to  $f(x) = x^5$  on the interval  $[0, 1]$  based on equally spaced data points. (b) Find the quadratic least squares approximation based on the data points  $0, .25, .5, .75, 1$ . (c) Find the quadratic least squares approximation with respect to the  $L^2$  norm. (d) Discuss the strengths and weaknesses of each approximation.
- 5.5.52. Let  $f(x) = x$ . Find the trigonometric function of the form  $g(x) = a + b \cos x + c \sin x$  that minimizes the  $L^2$  error  $\|g - f\| = \sqrt{\int_{-\pi}^{\pi} [g(x) - f(x)]^2 dx}$ .
- ◇ 5.5.53. Let  $g_1(t), \dots, g_n(t)$  be prescribed, linearly independent functions. Explain how to best approximate a function  $f(t)$  by a linear combination  $c_1 g_1(t) + \dots + c_n g_n(t)$  when the least squares error is measured in a weighted  $L^2$  norm  $\|f\|_w^2 = \int_a^b f(t)^2 w(t) dt$  with weight function  $w(t) > 0$ .
- 5.5.54. (a) Find the quadratic least squares approximation to  $f(t) = t^5$  on the interval  $[0, 1]$  with weights (i)  $w(t) = 1$ , (ii)  $w(t) = t$ , (iii)  $w(t) = e^{-t}$ . (b) Compare the errors — which gives the best result over the entire interval?

- 5.5.55. Let  $f_a(x) = \sqrt{\frac{a}{1+a^4x^2}}$ . Prove that (a)  $\|f_a\|_2 = \sqrt{\frac{\pi}{a}}$ , where  $\|\cdot\|_2$  denotes the  $L^2$  norm on  $(-\infty, \infty)$ . (b)  $\|f_a\|_\infty = \sqrt{a}$ , where  $\|\cdot\|_\infty$  denotes the  $L^\infty$  norm on  $(-\infty, \infty)$ . (c) Use this example to explain why having a small least squares error does not necessarily mean that the functions are everywhere close.

- 5.5.56. Find the plane  $z = \alpha + \beta x + \gamma y$  that best approximates the following functions on the square  $S = \{0 \leq x \leq 1, 0 \leq y \leq 1\}$  using the  $L^2$  norm  $\|f\|^2 = \iint_S |f(x, y)|^2 dx dy$  to measure the least squares error: (a)  $x^2 + y^2$ , (b)  $x^3 - y^3$ , (c)  $\sin \pi x \sin \pi y$ .

- 5.5.57. Find the radial polynomial  $p(x, y) = a + br + cr^2$ , where  $r^2 = x^2 + y^2$ , that best approximates the function  $f(x, y) = x$  using the  $L^2$  norm on the unit disk  $D = \{r \leq 1\}$  to measure the least squares error.

## Orthogonal Polynomials and Least Squares

In a similar fashion, the orthogonality of Legendre polynomials and their relatives serves to simplify the construction of least squares approximants in function space. Suppose, for instance, that our goal is to approximate the exponential function  $e^t$  by a polynomial on the interval  $-1 \leq t \leq 1$ , where the least squares error is measured using the standard  $L^2$  norm. We will write the best least squares approximant as a linear combination of the Legendre polynomials,

$$p(t) = a_0 P_0(t) + a_1 P_1(t) + \dots + a_n P_n(t) = a_0 + a_1 t + a_2 \left(\frac{3}{2}t^2 - \frac{1}{2}\right) + \dots \quad (5.78)$$

By orthogonality, the least squares coefficients can be immediately computed by the inner product formula (4.42), so, by the Rodrigues formula (4.61),

$$a_k = \frac{\langle e^t, P_k \rangle}{\|P_k\|^2} = \frac{2k+1}{2} \int_{-1}^1 e^t P_k(t) dt. \quad (5.79)$$

For example, the quadratic approximation is given by the first three terms in (5.78), whose coefficients are

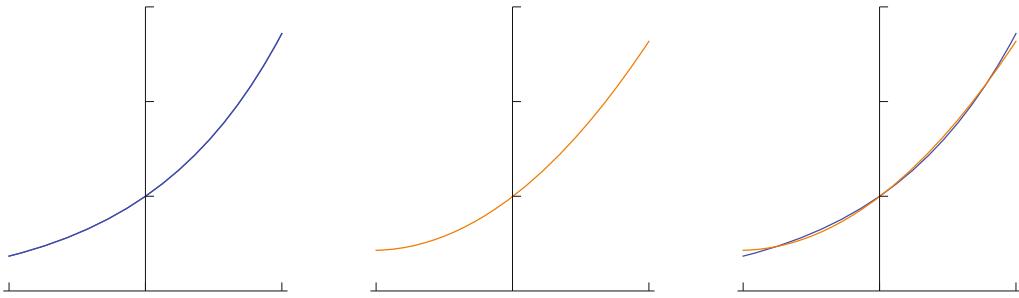
$$a_0 = \frac{1}{2} \int_{-1}^1 e^t dt = \frac{1}{2} \left(e - \frac{1}{e}\right) \simeq 1.175201, \quad a_1 = \frac{3}{2} \int_{-1}^1 t e^t dt = \frac{3}{e} \simeq 1.103638, \\ a_2 = \frac{5}{2} \int_{-1}^1 \left(\frac{3}{2}t^2 - \frac{1}{2}\right) e^t dt = \frac{5}{2} \left(e - \frac{7}{e}\right) \simeq .357814.$$

Graphs appear in [Figure 5.13](#); the first shows  $e^t$ , the second its quadratic approximant

$$e^t \approx 1.175201 + 1.103638 t + .357814 \left(\frac{3}{2}t^2 - \frac{1}{2}\right), \quad (5.80)$$

and the third compares the two by laying them on top of each other.

As in the discrete case, there are two major advantages of the orthogonal Legendre polynomials over the direct approach presented in Example 5.24. First, we do not need to solve any linear systems of equations since the required coefficients (5.79) are found by direct integration. Indeed, the coefficient matrix for polynomial least squares approximation based on the monomial basis is some variant of the notoriously ill-conditioned Hilbert matrix, (1.72), and the computation of an accurate solution can be tricky. Our precomputation of an orthogonal system of polynomials has successfully circumvented the ill-conditioned normal system.



**Figure 5.13.** Quadratic Least Squares Approximation to  $e^t$ .

The second advantage is that the coefficients  $a_k$  do *not* depend on the degree of the approximating polynomial. Thus, if the quadratic approximation (5.80) is insufficiently accurate, we merely append the cubic correction  $a_3 P_3(t)$  whose coefficient is given by

$$a_3 = \frac{7}{2} \int_{-1}^1 \left( \frac{5}{2} t^3 - \frac{3}{2} t \right) e^t dt = \frac{7}{2} \left( 37e - \frac{5}{e} \right) \simeq .070456,$$

Unlike the earlier method, there is no need to recompute the coefficients  $a_0, a_1, a_2$ , and hence the cubic least squares approximant is

$$e^t \approx 1.175201 + 1.103638 t + .357814 \left( \frac{3}{2} t^2 - \frac{1}{2} \right) + .070456 \left( \frac{5}{2} t^3 - \frac{3}{2} t \right). \quad (5.81)$$

And, if we desire yet further accuracy, we need only compute the next one or two coefficients.

To exploit orthogonality, each interval and norm requires the construction of a corresponding system of orthogonal polynomials. Let us reconsider Example 5.24, in which we used the method of least squares to approximate  $e^t$  based on the  $L^2$  norm on  $[0, 1]$ . Here, the ordinary Legendre polynomials are no longer orthogonal, and so we must use the rescaled Legendre polynomials (4.74) instead. Thus, the quadratic least squares approximant can be written as

$$\begin{aligned} p(t) &= a_0 + a_1 \tilde{P}_1(t) + a_2 \tilde{P}_2(t) = 1.718282 + .845155 (2t - 1) + .139864 (6t^2 - 6t + 1) \\ &= 1.012991 + .851125 t + .839184 t^2, \end{aligned}$$

where the coefficients

$$a_k = \frac{\langle e^t, \tilde{P}_k \rangle}{\| \tilde{P}_k \|^2} = (2k + 1) \int_0^1 \tilde{P}_k(t) e^t dt$$

are found by direct integration:

$$\begin{aligned} a_0 &= \int_0^1 e^t dt = e - 1 \simeq 1.718282, & a_1 &= 3 \int_0^1 (2t - 1) e^t dt = 3(3 - e) \simeq .845155, \\ a_2 &= 5 \int_0^1 (6t^2 - 6t + 1) e^t dt = 5(7e - 19) \simeq .139864. \end{aligned}$$

It is worth emphasizing that this is the *same* approximating polynomial we found earlier in (5.77). The use of an orthogonal system of polynomials merely streamlines the computation.

---

## Exercises

- 5.5.58. Use the Legendre polynomials to find the best (a) quadratic, and (b) cubic approximation to  $t^4$ , based on the  $L^2$  norm on  $[-1, 1]$ .
- 5.5.59. Repeat Exercise 5.5.58 using the  $L^2$  norm on  $[0, 1]$ .
- 5.5.60. Find the best cubic approximation to  $f(t) = e^t$  based on the  $L^2$  norm on  $[0, 1]$ .
- 5.5.61. Find the (a) linear, (b) quadratic, and (c) cubic polynomials  $q(t)$  that minimize the following integral:  $\int_0^1 [q(t) - t^3]^2 dt$ . What is the minimum value in each case?
- 5.5.62. Find the best quadratic and cubic approximations for  $\sin t$  for the  $L^2$  norm on  $[0, \pi]$  by using an orthogonal basis. Graph your results and estimate the maximal error.
- ♣ 5.5.63. Answer Exercise 5.5.60 when  $f(t) = \sin t$ . Use a computer to numerically evaluate the integrals.
- ♣ 5.5.64. Find the degree 6 least squares polynomial approximation to  $e^t$  on the interval  $[-1, 1]$  under the  $L^2$  norm.
- 5.5.65. (a) Use the polynomials and weighted norm from Exercise 4.5.12 to find the quadratic least squares approximation to  $f(t) = 1/t$ . In what sense is your quadratic approximation “best”? (b) Now find the best approximating cubic polynomial. (c) Compare the graphs of the quadratic and cubic approximants with the original function and discuss what you observe.
- ♣ 5.5.66. Use the Laguerre polynomials (4.68) to find the quadratic and cubic polynomial least squares approximation to  $f(t) = \tan^{-1} t$  relative to the weighted inner product (4.66). Use a computer to evaluate the coefficients. Graph your result and discuss what you observe.

---

## Splines

In pre-CAD (computer aided design) draftsmanship, a *spline* was a long, thin, flexible strip of wood or metal that was used to draw a smooth curve through prescribed points. The points were marked by small pegs, and the spline rested on the pegs. The mathematical theory of splines was first developed in the 1940s by the Romanian–American mathematician Isaac Schoenberg as an attractive alternative to polynomial interpolation and approximation. Splines have since become ubiquitous in numerical analysis, in geometric modeling, in design and manufacturing, in computer graphics and animation, and in many other applications.

We suppose that the spline coincides with the graph of a function  $y = u(x)$ . The pegs are fixed at the prescribed data points  $(x_0, y_0), \dots, (x_n, y_n)$ , and this requires  $u(x)$  to satisfy the interpolation conditions

$$u(x_j) = y_j, \quad j = 0, \dots, n. \quad (5.82)$$

The *mesh points*  $x_0 < x_1 < x_2 < \dots < x_n$  are distinct and labeled in increasing order. The spline is modeled as an elastic beam, and so satisfies the homogeneous beam equation  $u'''' = 0$ , cf. [61, 79]. Therefore,

$$u(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \quad \begin{array}{l} x_j \leq x \leq x_{j+1}, \\ j = 0, \dots, n-1, \end{array} \quad (5.83)$$

is a piecewise cubic function — meaning that, between successive mesh points, it is a cubic polynomial, but not necessarily the same cubic on each subinterval. The fact that we write the formula (5.83) in terms of  $x - x_j$  is merely for computational convenience.

Our problem is to determine the coefficients

$$a_j, \quad b_j, \quad c_j, \quad d_j, \quad j = 0, \dots, n - 1.$$

Since there are  $n$  subintervals, there is a total of  $4n$  coefficients, and so we require  $4n$  equations to uniquely prescribe them. First, we need the spline to satisfy the interpolation conditions (5.82). Since it is defined by a different formula on each side of the mesh point, this results in a total of  $2n$  conditions:

$$\begin{aligned} u(x_j^+) &= a_j = y_j, & j &= 0, \dots, n - 1, \\ u(x_{j+1}^-) &= a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 = y_{j+1}, \end{aligned} \quad (5.84)$$

where we abbreviate the length of the  $j^{\text{th}}$  subinterval by

$$h_j = x_{j+1} - x_j.$$

The next step is to require that the spline be as smooth as possible. The interpolation conditions (5.84) guarantee that  $u(x)$  is continuous. The condition that  $u(x) \in C^1$  be continuously differentiable requires that  $u'(x)$  be continuous at the interior mesh points  $x_1, \dots, x_{n-1}$ , which imposes the  $n - 1$  additional conditions

$$b_j + 2c_j h_j + 3d_j h_j^2 = u'(x_{j+1}^-) = u'(x_{j+1}^+) = b_{j+1}, \quad j = 0, \dots, n - 2. \quad (5.85)$$

To make  $u \in C^2$ , we impose  $n - 1$  further conditions

$$2c_j + 6d_j h_j = u''(x_{j+1}^-) = u''(x_{j+1}^+) = 2c_{j+1}, \quad j = 0, \dots, n - 2, \quad (5.86)$$

to ensure that  $u''$  is continuous at the mesh points. We have now imposed a total of  $4n - 2$  conditions, namely (5.84–86), on the  $4n$  coefficients. The two missing constraints will come from boundary conditions at the two endpoints, namely  $x_0$  and  $x_n$ . There are three common types:

(i) *Natural boundary conditions*:  $u''(x_0) = u''(x_n) = 0$ , whereby

$$c_0 = 0, \quad c_{n-1} + 3d_{n-1} h_{n-1} = 0. \quad (5.87)$$

Physically, this models a simply supported spline that rests freely on the first and last pegs.

(ii) *Clamped boundary conditions*:  $u'(x_0) = \alpha$ ,  $u'(x_n) = \beta$ , where  $\alpha, \beta$ , which could be 0, are specified by the user. This requires

$$b_0 = \alpha, \quad b_{n-1} + 2c_{n-1} h_{n-1} + 3d_{n-1} h_{n-1}^2 = \beta. \quad (5.88)$$

This corresponds to clamping the spline at prescribed angles at each end.

(iii) *Periodic boundary conditions*:  $u'(x_0) = u'(x_n)$ ,  $u''(x_0) = u''(x_n)$ , so that

$$b_0 = b_{n-1} + 2c_{n-1} h_{n-1} + 3d_{n-1} h_{n-1}^2, \quad c_0 = c_{n-1} + 3d_{n-1} h_{n-1}. \quad (5.89)$$

The periodic case is used to draw smooth closed curves; see below.

**Theorem 5.25.** Given mesh points  $a = x_0 < x_1 < \dots < x_n = b$ , and corresponding data values  $y_0, y_1, \dots, y_n$ , along with one of the three kinds of boundary conditions (5.87), (5.88), or (5.89), then there exists a unique piecewise cubic spline function  $u(x) \in C^2[a, b]$  that interpolates the data,  $u(x_0) = y_0, \dots, u(x_n) = y_n$ , and satisfies the boundary conditions.

*Proof:* We first discuss the natural case. The clamped case is left as an exercise for the reader, while the slightly harder periodic case will be treated at the end of the section. The first set of equations in (5.84) says that

$$a_j = y_j, \quad j = 0, \dots, n-1. \quad (5.90)$$

Next, (5.86–87) imply that

$$d_j = \frac{c_{j+1} - c_j}{3h_j}. \quad (5.91)$$

This equation also holds for  $j = n-1$ , provided that we make the convention that<sup>†</sup>

$$c_n = 0.$$

We now substitute (5.90–91) into the second set of equations in (5.84), and then solve the resulting equation for

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{(2c_j + c_{j+1})h_j}{3}. \quad (5.92)$$

Substituting this result and (5.91) back into (5.85), and simplifying, we obtain

$$h_j c_j + 2(h_j + h_{j+1})c_{j+1} + h_{j+1}c_{j+2} = 3 \left[ \frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j} \right] = z_{j+1}, \quad (5.93)$$

where we introduce  $z_{j+1}$  as a shorthand for the quantity on the right-hand side.

In the case of natural boundary conditions, we have

$$c_0 = 0, \quad c_n = 0,$$

and so (5.93) constitutes a tridiagonal linear system

$$A \mathbf{c} = \mathbf{z}, \quad (5.94)$$

for the unknown coefficients  $\mathbf{c} = (c_1, c_2, \dots, c_{n-1})^T$ , with coefficient matrix

$$A = \begin{pmatrix} 2(h_0 + h_1) & h_1 & & & & & \\ h_1 & 2(h_1 + h_2) & h_2 & & & & \\ & h_2 & 2(h_2 + h_3) & h_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & h_{n-3} & 2(h_{n-3} + h_{n-2}) & h_{n-2} & \\ & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & \end{pmatrix} \quad (5.95)$$

and right-hand side  $\mathbf{z} = (z_1, z_2, \dots, z_{n-1})^T$ . Once (5.95) has been solved, we will then use (5.90–92) to reconstruct the other spline coefficients  $a_j, b_j, d_j$ .

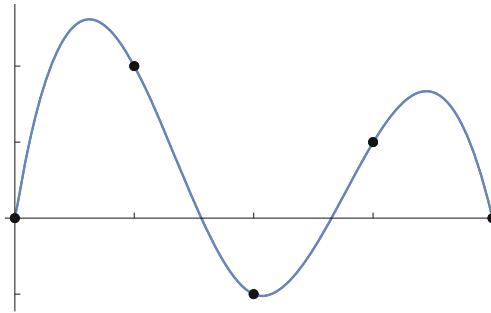
The key observation is that the coefficient matrix  $A$  is *strictly diagonally dominant*, meaning that each diagonal entry is strictly greater than the sum of the other entries in its row:

$$2(h_{j-1} + h_j) > h_{j-1} + h_j.$$

Theorem 8.19 below implies that  $A$  is nonsingular, and hence the tridiagonal linear system has a unique solution  $\mathbf{c}$ . This suffices to prove the theorem in the case of natural boundary conditions. *Q.E.D.*

---

<sup>†</sup> This is merely for convenience; there is no  $c_n$  used in the formula for the spline.



**Figure 5.14.** A Cubic Spline.

To actually solve the linear system (5.94), we can apply our tridiagonal solution algorithm (1.68). Let us specialize to the most important case, in which the mesh points are equally spaced in the interval  $[a, b]$ , so that

$$x_j = a + j h, \quad \text{where} \quad h = h_j = \frac{b - a}{n}, \quad j = 0, \dots, n - 1.$$

In this case, the coefficient matrix  $A = h B$  is equal to  $h$  times the tridiagonal matrix

$$B = \begin{pmatrix} 4 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & 1 & 4 & 1 & \\ & & & 1 & 4 & 1 \\ & & & & \ddots & \ddots \end{pmatrix}$$

that first appeared in Example 1.37. Its  $LU$  factorization takes on a rather simple form, since most of the entries of  $L$  and  $U$  are essentially the same, modulo rounding error. This makes the implementation of the Forward and Back Substitution procedures almost trivial.

Figure 5.14 shows a particular example — a natural spline passing through the data points  $(0, 0), (1, 2), (2, -1), (3, 1), (4, 0)$ . The human eye is unable to discern the discontinuities in its third derivatives, and so the graph appears completely smooth, even though it is, in fact, only  $C^2$ .

In the periodic case, we set

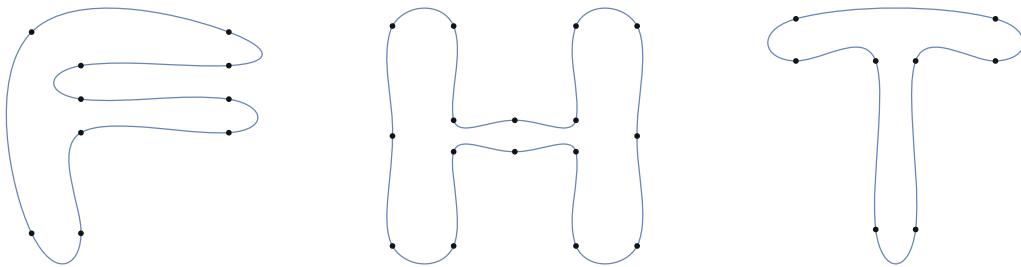
$$a_{n+k} = a_n, \quad b_{n+k} = b_n, \quad c_{n+k} = c_n, \quad d_{n+k} = d_n, \quad z_{n+k} = z_n.$$

With this convention, the basic equations (5.90–93) are the same. In this case, the coefficient matrix for the linear system

$$A \mathbf{c} = \mathbf{z}, \quad \text{with} \quad \mathbf{c} = (c_0, c_1, \dots, c_{n-1})^T, \quad \mathbf{z} = (z_0, z_1, \dots, z_{n-1})^T,$$

is of *tricirculant* form:

$$A = \begin{pmatrix} 2(h_{n-1} + h_0) & h_0 & & & & h_{n-1} \\ h_0 & 2(h_0 + h_1) & h_1 & & & \\ & h_1 & 2(h_1 + h_2) & h_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & h_{n-3} & 2(h_{n-3} + h_{n-2}) & h_{n-2} \\ & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix}. \quad (5.96)$$



**Figure 5.15.** Three Sample Spline Letters.

Again  $A$  is strictly diagonally dominant, and so there is a unique solution  $\mathbf{c}$ , from which one reconstructs the spline, proving Theorem 5.25 in the periodic case. The  $LU$  factorization of tricirculant matrices was discussed in Exercise 1.7.14.

One immediate application of splines is curve fitting in computer aided design and graphics. The basic problem is to draw a smooth parameterized curve  $\mathbf{u}(t) = (u(t), v(t))^T$  that passes through a set of prescribed data points  $\mathbf{x}_k = (x_k, y_k)^T$  in the plane. We have the freedom to choose the parameter value  $t = t_k$  when the curve passes through the  $k^{\text{th}}$  point; the simplest and most common choice is to set  $t_k = k$ . We then construct the functions  $x = u(t)$  and  $y = v(t)$  as cubic splines interpolating the  $x$  and  $y$  coordinates of the data points, so  $u(t_k) = x_k$ ,  $v(t_k) = y_k$ . For smooth closed curves, we require that both splines be periodic; for curves with ends, either natural or clamped boundary conditions are used.

Most computer graphics packages include one or more implementations of parameterized spline curves. The same idea also underlies modern font design for laser printing and typography (including the fonts used in this book). The great advantage of spline fonts over their bitmapped counterparts is that they can be readily scaled. Some sample letter shapes parameterized by periodic splines passing through the indicated data points are plotted in Figure 5.15. Better fits can be easily obtained by increasing the number of data points. Various extensions of the basic spline algorithms to space curves and surfaces are an essential component of modern computer graphics, design, and animation, [25, 74].

## Exercises

5.5.67. Find and graph the natural cubic spline interpolant for the following data:

$$(a) \begin{array}{c|ccc} x & -1 & 0 & 1 \\ \hline y & -2 & 1 & -1 \end{array}$$

$$(b) \begin{array}{c|ccccc} x & 0 & 1 & 2 & 3 \\ \hline y & 1 & 2 & 0 & 1 \end{array}$$

$$(c) \begin{array}{c|ccc} x & 1 & 2 & 4 \\ \hline y & 3 & 0 & 2 \end{array}$$

$$(d) \begin{array}{c|cccccc} x & -2 & -1 & 0 & 1 & 2 \\ \hline y & 5 & 2 & 3 & -1 & 1 \end{array}$$

5.5.68. Repeat Exercise 5.5.67 when the spline has homogeneous clamped boundary conditions.

5.5.69. Find and graph the periodic cubic spline that interpolates the following data:

(a)	$\begin{array}{c cccc} x & 0 & 1 & 2 & 3 \\ \hline y & 1 & 0 & 0 & 1 \end{array}$
-----	---

(b)	$\begin{array}{c cccc} x & 0 & 1 & 2 & 3 \\ \hline y & 1 & 2 & 0 & 1 \end{array}$
-----	---

(c)	$\begin{array}{c ccccc} x & 0 & 1 & 2 & 3 & 4 \\ \hline y & 1 & 0 & 0 & 0 & 1 \end{array}$
-----	--

(d)	$\begin{array}{c cccccc} x & -2 & -1 & 0 & 1 & 2 \\ \hline y & 1 & 2 & -2 & -1 & 1 \end{array}$
-----	---

♠ 5.5.70. (a) Given the known values of  $\sin x$  at  $x = 0^\circ, 30^\circ, 45^\circ, 60^\circ$ , construct the natural cubic spline interpolant. (b) Compare the accuracy of the spline with the least squares and interpolating polynomials you found in Exercise 5.5.21.

♣ 5.5.71. (a) Using the exact values for  $\sqrt{x}$  at  $x = 0, \frac{1}{4}, \frac{9}{16}, 1$ , construct the natural cubic spline interpolant. (b) What is the maximal error of the spline on the interval  $[0, 1]$ ? (c) Compare the error with that of the interpolating cubic polynomial you found in Exercise 5.5.23. Which is the better approximation? (d) Answer part (d) using the cubic least squares approximant based on the  $L^2$  norm on  $[0, 1]$ .

♠ 5.5.72. According to Figure 5.9, the interpolating polynomials for the function  $1/(1+x^2)$  on the interval  $[-3, 3]$  based on equally spaced mesh points are very inaccurate near the ends of the interval. Does the natural spline interpolant based on the same 3, 5, and 11 data points exhibit the same inaccuracy?

♣ 5.5.73. (a) Draw outlines of the block capital letters I, C, S, and Y on a sheet of graph paper. Fix several points on the graphs and measure their  $x$  and  $y$  coordinates. (b) Use periodic cubic splines  $x = u(t), y = v(t)$  to interpolate the coordinates of the data points using equally spaced nodes for the parameter values  $t_k$ . Graph the resulting spline letters, and discuss how the method could be used in font design. To get nicer results, you may wish to experiment with different numbers and locations for the points.

♣ 5.5.74. Repeat Exercise 5.5.73, using the Lagrange interpolating polynomials instead of splines to parameterize the curves. Compare the two methods and discuss advantages and disadvantages.

♡ 5.5.75. Let  $x_0 < x_1 < \dots < x_n$ . For each  $j = 0, \dots, n$ , the  $j^{\text{th}}$  cardinal spline  $C_j(x)$  is defined to be the natural cubic spline interpolating the Lagrange data

$$y_0 = 0, \quad y_1 = 0, \quad \dots \quad y_{j-1} = 0, \quad y_j = 1, \quad y_{j+1} = 0, \quad \dots \quad y_n = 0.$$

(a) Construct and graph the natural cardinal splines corresponding to the nodes  $x_0 = 0$ ,  $x_1 = 1$ ,  $x_2 = 2$ , and  $x_3 = 3$ . (b) Prove that the natural spline that interpolates the data  $y_0, \dots, y_n$  can be uniquely written as a linear combination  $u(x) = y_0 C_0(x) + y_1 C_1(x) + \dots + y_n C_n(x)$  of the cardinal splines. (c) Explain why the space of natural splines on  $n+1$  nodes is a vector space of dimension  $n+1$ . (d) Discuss briefly what modifications are required to adapt this method to periodic and to clamped splines.

♡ 5.5.76. A bell-shaped or B-spline  $u = \beta(x)$  interpolates the data

$$\beta(-2) = 0, \quad \beta(-1) = 1, \quad \beta(0) = 4, \quad \beta(1) = 1, \quad \beta(2) = 0.$$

(a) Find the explicit formula for the natural B-spline and plot its graph. (b) Show that  $\beta(x)$  also satisfies the homogeneous clamped boundary conditions  $u'(-2) = u'(2) = 0$ . (c) Show that  $\beta(x)$  also satisfies the periodic boundary conditions. Thus, for this particular interpolation problem, the natural, clamped, and periodic splines happen to coincide. (d) Show that  $\beta^*(x) = \begin{cases} \beta(x), & -2 \leq x \leq 2 \\ 0, & \text{otherwise,} \end{cases}$  defines a  $C^2$  spline on every interval  $[-k, k]$ .

♡ 5.5.77. Let  $\beta(x)$  denote the B-spline function of Exercise 5.5.76. Assuming  $n \geq 4$ , let  $P_n$  denote the vector space of periodic cubic splines based on the integer nodes  $x_j = j$  for  $j = 0, \dots, n$ . (a) Prove that the B-splines  $B_j(x) = \beta((x - j - m) \bmod n + m)$ ,

$j = 0, \dots, n - 1$ , where  $m$  denotes the integer part of  $n/2$ , form a basis for  $P_n$ . (b) Graph the basis periodic  $B$ -splines in the case  $n = 5$ . (c) Let  $u(x)$  denote the periodic spline interpolant for the data values  $y_0, \dots, y_{n-1}$ . Explain how to write  $u(x) = \alpha_0 B_0(x) + \dots + \alpha_{n-1} B_{n-1}(x)$  in terms of the  $B$ -splines by solving a linear system for the coefficients  $\alpha_0, \dots, \alpha_{n-1}$ . (d) Write the periodic spline with  $y_0 = y_5 = 0$ ,  $y_1 = 2$ ,  $y_2 = 1$ ,  $y_3 = -1$ ,  $y_4 = -2$ , as a linear combination of the periodic basis  $B$ -splines  $B_0(x), \dots, B_4(x)$ . Plot the resulting periodic spline function.

## 5.6 Discrete Fourier Analysis and the Fast Fourier Transform

In modern digital media — audio, still images, video, etc. — continuous signals are sampled at discrete time intervals before being processed. Fourier analysis decomposes the sampled signal into its fundamental periodic constituents — sines and cosines, or, more conveniently, complex exponentials. The crucial fact, upon which all of modern signal processing is based, is that the sampled complex exponentials form an orthogonal basis. This section introduces the Discrete Fourier Transform, and concludes with an introduction to the justly famous Fast Fourier Transform, an efficient algorithm for computing the discrete Fourier representation and reconstructing the signal from its Fourier coefficients.

We will concentrate on the one-dimensional version here. Let  $f(x)$  be a function representing the signal, defined on an interval  $a \leq x \leq b$ . Our computer can store its measured values only at a finite number of *sample points*  $a \leq x_0 < x_1 < \dots < x_n \leq b$ . In the simplest, and by far the most common, case, the sample points are equally spaced, and so

$$x_j = a + j h, \quad j = 0, \dots, n, \quad \text{where} \quad h = \frac{b - a}{n}$$

indicates the sample rate. In signal processing applications,  $x$  represents time instead of space, and the  $x_j$  are the times at which we sample the signal  $f(x)$ . Sample rates can be very high, e.g., every 10–20 milliseconds in current speech recognition systems.

For simplicity, we adopt the “standard” interval of  $0 \leq x \leq 2\pi$ , and the  $n$  equally spaced sample points<sup>†</sup>

$$x_0 = 0, \quad x_1 = \frac{2\pi}{n}, \quad x_2 = \frac{4\pi}{n}, \quad \dots \quad x_j = \frac{2j\pi}{n}, \quad \dots \quad x_{n-1} = \frac{2(n-1)\pi}{n}. \quad (5.97)$$

(Signals defined on other intervals can be handled by simply rescaling the interval to have length  $2\pi$ .) Sampling a (complex-valued) signal or function  $f(x)$  produces the *sample vector*

$$\mathbf{f} = (f_0, f_1, \dots, f_{n-1})^T = (f(x_0), f(x_1), \dots, f(x_{n-1}))^T,$$

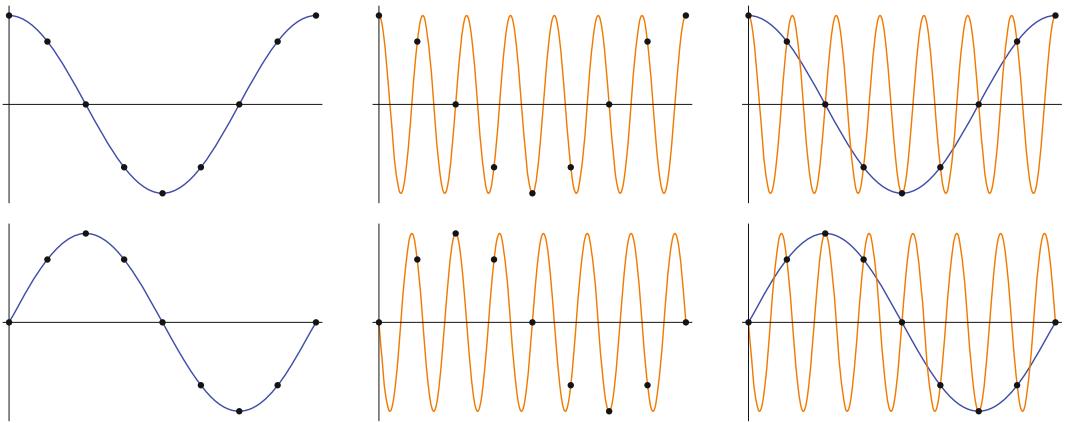
where

$$f_j = f(x_j) = f\left(\frac{2j\pi}{n}\right), \quad j = 0, \dots, n - 1. \quad (5.98)$$

Sampling cannot distinguish between functions that have the same values at all of the sample points — from the sampler’s point of view they are identical. For example, the periodic complex exponential function

$$f(x) = e^{inx} = \cos nx + i \sin nx$$

<sup>†</sup> We will find it convenient to omit the final sample point  $x_n = 2\pi$  from consideration.



**Figure 5.16.** Sampling  $e^{-ix}$  and  $e^{7ix}$  on  $n = 8$  sample points.

has sampled values

$$f_j = f\left(\frac{2j\pi}{n}\right) = \exp\left(i n \frac{2j\pi}{n}\right) = e^{2j\pi i} = 1 \quad \text{for all } j = 0, \dots, n-1,$$

and hence is indistinguishable from the constant function  $c(x) \equiv 1$  — both lead to the *same* sample vector  $(1, 1, \dots, 1)^T$ . This has the important implication that sampling at  $n$  equally spaced sample points *cannot* detect periodic signals of frequency  $n$ . More generally, the two complex exponential signals

$$e^{i(k+n)x} \quad \text{and} \quad e^{ikx}$$

are also indistinguishable when sampled. Consequently, we need only use the first  $n$  periodic complex exponential functions

$$f_0(x) = 1, \quad f_1(x) = e^{ix}, \quad f_2(x) = e^{2ix}, \quad \dots \quad f_{n-1}(x) = e^{(n-1)ix}, \quad (5.99)$$

in order to represent any  $2\pi$  periodic sampled signal. In particular, exponentials  $e^{-ikx}$  of “negative” frequency can all be converted into positive versions, namely  $e^{i(n-k)x}$ , by the same sampling argument. For example,

$$e^{-ix} = \cos x - i \sin x \quad \text{and} \quad e^{(n-1)ix} = \cos(n-1)x + i \sin(n-1)x$$

have identical values on the sample points (5.97). However, off of the sample points, they are quite different; the former is slowly varying, while the latter represents a high-frequency oscillation. In Figure 5.16, we compare  $e^{-ix}$  and  $e^{7ix}$  when there are  $n = 8$  sample values, indicated by the dots on the graphs. The top row compares the real parts,  $\cos x$  and  $\cos 7x$ , while the bottom row compares the imaginary parts,  $\sin x$  and  $-\sin 7x$ . Note that both functions have the same pattern of sample values, even though their overall behavior is strikingly different.

This effect is commonly referred to as *aliasing*.<sup>†</sup> If you view a moving particle under a stroboscopic light that flashes only eight times, you would be unable to determine which

---

<sup>†</sup> In computer graphics, the term “aliasing” is used in a much broader sense that covers a variety of artifacts introduced by discretization — particularly, the jagged appearance of lines and smooth curves on a digital monitor.

of the two graphs the particle was following. Aliasing is the cause of a well-known artifact in movies: spoked wheels can appear to be rotating backwards when our brain interprets the discretization of the high-frequency forward motion imposed by the frames of the film as an equivalently discretized low-frequency motion in reverse. Aliasing also has important implications for the design of music CD's. We must sample an audio signal at a sufficiently high rate that all audible frequencies can be adequately represented. In fact, human appreciation of music also relies on inaudible high-frequency tones, and so a much higher sample rate is actually used in commercial CD design. But the sample rate that was selected remains controversial; hi fi aficionados complain that it was not set high enough to fully reproduce the musical quality of an analog LP record!

The *discrete Fourier representation* decomposes a sampled function  $f(x)$  into a linear combination of complex exponentials. Since we cannot distinguish sampled exponentials of frequency higher than  $n$ , we only need consider a finite linear combination

$$f(x) \sim p(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + \cdots + c_{n-1} e^{(n-1)ix} = \sum_{k=0}^{n-1} c_k e^{ikx} \quad (5.100)$$

of the first  $n$  exponentials (5.99). The symbol  $\sim$  in (5.100) means that the function  $f(x)$  and the sum  $p(x)$  agree on the sample points:

$$f(x_j) = p(x_j), \quad j = 0, \dots, n-1. \quad (5.101)$$

Therefore,  $p(x)$  can be viewed as a (complex-valued) *interpolating trigonometric polynomial* of degree  $\leq n-1$  for the sample data  $f_j = f(x_j)$ .

**Remark.** If  $f(x)$  is real, then  $p(x)$  is also real on the sample points, but may very well be complex-valued in between. To avoid this unsatisfying state of affairs, we will usually discard its imaginary component, and regard the real part of  $p(x)$  as “the” interpolating trigonometric polynomial. On the other hand, sticking with a purely real construction unnecessarily complicates the underlying mathematical analysis, and so we will retain the complex exponential form (5.100) of the discrete Fourier sum.

Since we are working in the finite-dimensional complex vector space  $\mathbb{C}^n$  throughout, we can reformulate the discrete Fourier series in vectorial form. Sampling the basic exponentials (5.99) produces the complex vectors

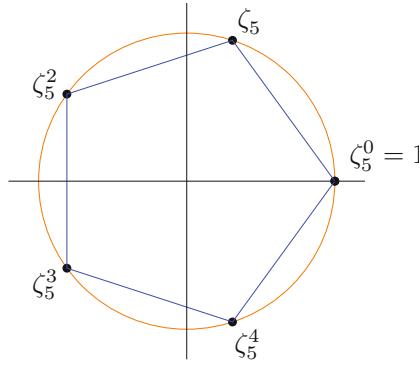
$$\begin{aligned} \boldsymbol{\omega}_k &= (e^{ikx_0}, e^{ikx_1}, e^{ikx_2}, \dots, e^{ikx_{n-1}})^T \\ &= \left(1, e^{2k\pi i/n}, e^{4k\pi i/n}, \dots, e^{2(n-1)k\pi i/n}\right)^T, \end{aligned} \quad k = 0, \dots, n-1. \quad (5.102)$$

The interpolation conditions (5.101) can be recast in the equivalent vector form

$$\mathbf{f} = c_0 \boldsymbol{\omega}_0 + c_1 \boldsymbol{\omega}_1 + \cdots + c_{n-1} \boldsymbol{\omega}_{n-1}. \quad (5.103)$$

In other words, to compute the discrete Fourier coefficients  $c_0, \dots, c_{n-1}$  of  $f$ , all we need to do is rewrite its sample vector  $\mathbf{f}$  as a linear combination of the sampled exponential vectors  $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$ .

Now, the absolutely crucial property is the orthonormality of the basis elements  $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$ . Were it not for the power of orthogonality, Fourier analysis might have remained a mere mathematical curiosity, rather than today's indispensable tool.



**Figure 5.17.** The Fifth Roots of Unity.

**Proposition 5.26.** The sampled exponential vectors  $\omega_0, \dots, \omega_{n-1}$  form an orthonormal basis of  $\mathbb{C}^n$  with respect to the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle = \frac{1}{n} \sum_{j=0}^{n-1} f_j \overline{g_j} = \frac{1}{n} \sum_{j=0}^{n-1} f(x_j) \overline{g(x_j)}, \quad \mathbf{f}, \mathbf{g} \in \mathbb{C}^n. \quad (5.104)$$

**Remark.** The inner product (5.104) is a rescaled version of the standard Hermitian dot product (3.98) between complex vectors. We can interpret the inner product between the sample vectors  $\mathbf{f}, \mathbf{g}$  as the *average* of the sampled values of the product signal  $f(x) \overline{g(x)}$ .

*Proof:* The crux of the matter relies on properties of the remarkable complex numbers

$$\zeta_n = e^{2\pi i/n} = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n}, \quad \text{where } n = 1, 2, 3, \dots. \quad (5.105)$$

Particular cases include

$$\zeta_2 = -1, \quad \zeta_3 = -\frac{1}{2} + \frac{\sqrt{3}}{2} i, \quad \zeta_4 = i, \quad \text{and} \quad \zeta_8 = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} i. \quad (5.106)$$

The  $n^{\text{th}}$  power of  $\zeta_n$  is

$$\zeta_n^n = \left( e^{2\pi i/n} \right)^n = e^{2\pi i} = 1,$$

and hence  $\zeta_n$  is one of the complex  $n^{\text{th}}$  roots of unity:  $\zeta_n = \sqrt[n]{1}$ . There are, in fact,  $n$  distinct complex  $n^{\text{th}}$  roots of 1, including 1 itself, namely the powers of  $\zeta_n$ :

$$\zeta_n^k = e^{2k\pi i/n} = \cos \frac{2k\pi}{n} + i \sin \frac{2k\pi}{n}, \quad k = 0, \dots, n-1. \quad (5.107)$$

Since it generates all the others,  $\zeta_n$  is known as a *primitive  $n^{\text{th}}$  root of unity*. Geometrically, the  $n^{\text{th}}$  roots (5.107) are the vertices of a regular unit  $n$ -gon inscribed in the unit circle  $|z| = 1$ ; see Figure 5.17 for the case  $n = 5$ , where the roots form the vertices of a regular pentagon. The primitive root  $\zeta_n$  is the first vertex we encounter as we go around the  $n$ -gon in a counterclockwise direction, starting at 1. Continuing around, the other roots appear in their natural order  $\zeta_n^2, \zeta_n^3, \dots, \zeta_n^{n-1}$ , cycling back to  $\zeta_n^n = 1$ . The complex conjugate of  $\zeta_n$  is the “last”  $n^{\text{th}}$  root:

$$e^{-2\pi i/n} = \overline{\zeta_n} = \frac{1}{\zeta_n} = \zeta_n^{n-1} = e^{2(n-1)\pi i/n}. \quad (5.108)$$

The complex numbers (5.107) are a complete set of roots of the polynomial  $z^n - 1$ , which can therefore be completely factored:

$$z^n - 1 = (z - 1)(z - \zeta_n)(z - \zeta_n^2) \cdots (z - \zeta_n^{n-1}).$$

On the other hand, elementary algebra provides us with the real factorization

$$z^n - 1 = (z - 1)(1 + z + z^2 + \cdots + z^{n-1}).$$

Comparing the two, we conclude that

$$1 + z + z^2 + \cdots + z^{n-1} = (z - \zeta_n)(z - \zeta_n^2) \cdots (z - \zeta_n^{n-1}).$$

Substituting  $z = \zeta_n^k$  into both sides of this identity, we deduce the useful formula

$$1 + \zeta_n^k + \zeta_n^{2k} + \cdots + \zeta_n^{(n-1)k} = \begin{cases} n, & k = 0, \\ 0, & 0 < k < n. \end{cases} \quad (5.109)$$

Since  $\zeta_n^{n+k} = \zeta_n^k$ , this formula can easily be extended to general integers  $k$ ; the sum is equal to  $n$  if  $n$  evenly divides  $k$  and is 0 otherwise.

Now, let us apply what we've learned to prove Proposition 5.26. First, in view of (5.107), the sampled exponential vectors (5.102) can all be written in terms of the  $n^{\text{th}}$  roots of unity:

$$\boldsymbol{\omega}_k = (1, \zeta_n^k, \zeta_n^{2k}, \zeta_n^{3k}, \dots, \zeta_n^{(n-1)k})^T, \quad k = 0, \dots, n-1. \quad (5.110)$$

Therefore, applying (5.108, 109), we conclude that

$$\langle \boldsymbol{\omega}_k, \boldsymbol{\omega}_l \rangle = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{jk} \overline{\zeta_n^{jl}} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{j(k-l)} = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \quad 0 \leq k, l < n,$$

which establishes orthonormality of the sampled exponential vectors. Q.E.D.

Orthonormality of the basis vectors implies that we can immediately compute the Fourier coefficients in the discrete Fourier sum (5.100) by taking inner products:

$$c_k = \langle \mathbf{f}, \boldsymbol{\omega}_k \rangle = \frac{1}{n} \sum_{j=0}^{n-1} f_j \overline{e^{ikx_j}} = \frac{1}{n} \sum_{j=0}^{n-1} f_j e^{-ikx_j} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{-jk} f_j. \quad (5.111)$$

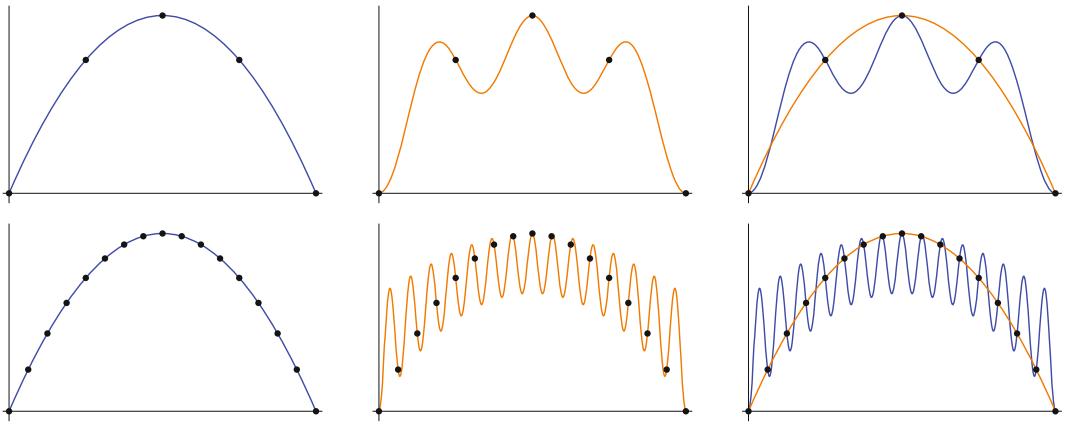
In other words, the discrete Fourier coefficient  $c_k$  is obtained by averaging the sampled values of the product function  $f(x) e^{-ikx}$ . The passage from a signal to its Fourier coefficients is known as the *Discrete Fourier Transform* or DFT for short. The reverse procedure of reconstructing a signal from its discrete Fourier coefficients via the sum (5.100) (or (5.103)) is known as the *Inverse Discrete Fourier Transform* or IDFT.

**Example 5.27.** If  $n = 4$ , then  $\zeta_4 = i$ . The corresponding sampled exponential vectors

$$\boldsymbol{\omega}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{\omega}_1 = \begin{pmatrix} 1 \\ i \\ -1 \\ -i \end{pmatrix}, \quad \boldsymbol{\omega}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \quad \boldsymbol{\omega}_3 = \begin{pmatrix} 1 \\ -i \\ -1 \\ i \end{pmatrix},$$

form an orthonormal basis of  $\mathbb{C}^4$  with respect to the averaged Hermitian dot product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \frac{1}{4} (v_0 \overline{w_0} + v_1 \overline{w_1} + v_2 \overline{w_2} + v_3 \overline{w_3}), \quad \text{where} \quad \mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix}.$$



**Figure 5.18.** The Discrete Fourier Representation of  $2\pi x - x^2$ .

Given the sampled function values

$$f_0 = f(0), \quad f_1 = f\left(\frac{1}{2}\pi\right), \quad f_2 = f(\pi), \quad f_3 = f\left(\frac{3}{2}\pi\right),$$

we construct the discrete Fourier representation

$$\mathbf{f} = c_0 \omega_0 + c_1 \omega_1 + c_2 \omega_2 + c_3 \omega_3, \quad (5.112)$$

where

$$\begin{aligned} c_0 &= \langle \mathbf{f}, \omega_0 \rangle = \frac{1}{4}(f_0 + f_1 + f_2 + f_3), & c_1 &= \langle \mathbf{f}, \omega_1 \rangle = \frac{1}{4}(f_0 - if_1 - f_2 + if_3), \\ c_2 &= \langle \mathbf{f}, \omega_2 \rangle = \frac{1}{4}(f_0 - f_1 + f_2 - f_3), & c_3 &= \langle \mathbf{f}, \omega_3 \rangle = \frac{1}{4}(f_0 + if_1 - f_2 - if_3). \end{aligned}$$

We interpret this decomposition as the complex exponential interpolant

$$f(x) \sim p(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + c_3 e^{3ix}$$

that agrees with  $f(x)$  on the 4 sample points.

For instance, if

$$f(x) = 2\pi x - x^2,$$

then

$$f_0 = 0, \quad f_1 = 7.4022, \quad f_2 = 9.8696, \quad f_3 = 7.4022,$$

and hence

$$c_0 = 6.1685, \quad c_1 = -2.4674, \quad c_2 = -1.2337, \quad c_3 = -2.4674.$$

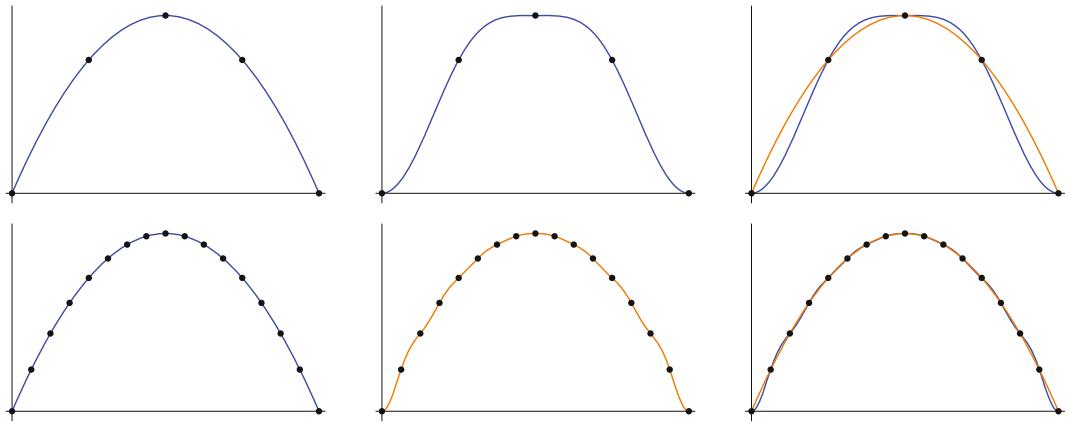
Therefore, the interpolating trigonometric polynomial is given by the real part of

$$p_4(x) = 6.1685 - 2.4674 e^{ix} - 1.2337 e^{2ix} - 2.4674 e^{3ix}, \quad (5.113)$$

namely,

$$\operatorname{Re} p_4(x) = 6.1685 - 2.4674 \cos x - 1.2337 \cos 2x - 2.4674 \cos 3x. \quad (5.114)$$

In [Figure 5.18](#), we compare the function, with the interpolation points indicated, and its discrete Fourier representations (5.114) for both  $n = 4$  in the first row, and  $n = 16$  points in the second. The resulting graphs point out a significant difficulty with the Discrete Fourier Transform as developed so far. While the trigonometric polynomials do indeed correctly match the sampled function values, their pronounced oscillatory behavior makes them completely unsuitable for interpolation away from the sample points.



**Figure 5.19.** The Low-Frequency Discrete Fourier Representation of  $x^2 - 2\pi x$ .

However, this difficulty can be rectified by being a little more clever. The problem is that we have not been paying sufficient attention to the frequencies that are represented in the Fourier sum. Indeed, the graphs in Figure 5.18 might remind you of our earlier observation that, due to aliasing, low- and high-frequency exponentials can have the same sample data, but differ wildly in between the sample points. While the first half of the summands in (5.100) represent relatively low frequencies, the second half do not, and can be replaced by equivalent lower-frequency, and hence less-oscillatory, exponentials. Namely, if  $0 < k \leq \frac{1}{2}n$ , then  $e^{-i k x}$  and  $e^{i(n-k)x}$  have the same sample values, but the former is of lower frequency than the latter. Thus, for interpolatory purposes, we should replace the second half of the summands in the Fourier sum (5.100) by their low-frequency alternatives. If  $n = 2m + 1$  is odd, then we take

$$\hat{p}_{2m+1}(x) = c_{-m} e^{-i m x} + \cdots + c_{-1} e^{-i x} + c_0 + c_1 e^{i x} + \cdots + c_m e^{i m x} = \sum_{k=-m}^m c_k e^{i k x} \quad (5.115)$$

as the equivalent low-frequency interpolant. If  $n = 2m$  is even — which is the most common case occurring in applications — then

$$\hat{p}_{2m}(x) = c_{-m} e^{-i m x} + \cdots + c_{-1} e^{-i x} + c_0 + c_1 e^{i x} + \cdots + c_{m-1} e^{i(m-1)x} = \sum_{k=-m}^{m-1} c_k e^{i k x} \quad (5.116)$$

will be our choice. (It is a matter of personal taste whether to use  $e^{-i m x}$  or  $e^{i m x}$  to represent the highest-frequency term.) In both cases, the Fourier coefficients with negative indices are the same as their high-frequency alternatives:

$$c_{-k} = c_{n-k} = \langle \mathbf{f}, \boldsymbol{\omega}_{n-k} \rangle = \langle \mathbf{f}, \boldsymbol{\omega}_{-k} \rangle, \quad (5.117)$$

where  $\boldsymbol{\omega}_{-k} = \boldsymbol{\omega}_{n-k}$  is the sample vector for  $e^{-i k x} \sim e^{i(n-k)x}$ .

Returning to the previous example, for interpolating purposes, we should replace (5.113) by the equivalent low-frequency interpolant

$$\hat{p}_4(x) = -1.2337 e^{-2i x} - 2.4674 e^{-i x} + 6.1685 - 2.4674 e^{i x}, \quad (5.118)$$

with real part

$$\operatorname{Re} \hat{p}_4(x) = 6.1685 - 4.9348 \cos x - 1.2337 \cos 2x.$$

Graphs of the  $n = 4$  and 16 low-frequency trigonometric interpolants can be seen in [Figure 5.19](#). Thus, by utilizing only the lowest-frequency exponentials, we successfully suppress the aliasing artifacts, resulting in a quite reasonable trigonometric interpolant to the given function on the entire interval.

**Remark.** The low-frequency version also serves to unravel the reality of the Fourier representation of a real function  $f(x)$ . Since  $\omega_{-k} = \overline{\omega_k}$ , formula (5.117) implies that  $c_{-k} = \overline{c_k}$ , and so the common frequency terms

$$c_{-k} e^{-ikx} + c_k e^{ikx} = a_k \cos kx + b_k \sin kx$$

add up to a real trigonometric function. Therefore, the odd  $n$  interpolant (5.115) is a real trigonometric polynomial, whereas in the even version (5.116) only the highest-frequency term  $c_{-m} e^{-imx}$  produces a complex term — which is, in fact, 0 on the sample points.

## Exercises

5.6.1. Find (i) the discrete Fourier coefficients, and (ii) the low-frequency trigonometric interpolant, for the following functions using the indicated number of sample points: (a)  $\sin x$ ,

$$n = 4, \quad (b) |x - \pi|, \quad n = 6, \quad (c) f(x) = \begin{cases} 1, & x \leq 2, \\ 0, & x > 2, \end{cases} \quad n = 6, \quad (d) \operatorname{sign}(x - \pi), \quad n = 8.$$

5.6.2. Find (i) the sample values, and (ii) the trigonometric interpolant corresponding to the following discrete Fourier coefficients: (a)  $c_{-1} = c_1 = 1, c_0 = 0$ ,

$$(b) c_{-2} = c_0 = c_2 = 1, c_{-1} = c_1 = -1, \quad (c) c_{-2} = c_0 = c_1 = 2, c_{-1} = c_2 = 0, \\ (d) c_0 = c_2 = c_4 = 1, c_1 = c_3 = c_5 = -1.$$

♣ 5.6.3. Let  $f(x) = x$ . Compute its discrete Fourier coefficients based on  $n = 4, 8$  and 16 interpolation points. Then, plot  $f(x)$  along with the resulting (real) trigonometric interpolants and discuss their accuracy.

♣ 5.6.4. Answer Exercise 5.6.3 for the functions (a)  $x^2$ , (b)  $(x - \pi)^2$ , (c)  $\sin x$ , (d)  $\cos \frac{1}{2}x$ , (e)  $\begin{cases} 1, & \frac{1}{2}\pi \leq x \leq \frac{3}{2}\pi, \\ 0, & \text{otherwise,} \end{cases}$  (f)  $\begin{cases} x, & 0 \leq x \leq \pi, \\ 2\pi - x, & \pi \leq x \leq 2\pi. \end{cases}$

5.6.5. (a) Draw a picture of the complex plane with the complex solutions to  $z^6 = 1$  marked.

(b) What is the exact formula (no trigonometric functions allowed) for the primitive sixth root of unity  $\zeta_6$ ? (c) Verify explicitly that  $1 + \zeta_6 + \zeta_6^2 + \zeta_6^3 + \zeta_6^4 + \zeta_6^5 = 0$ . (d) Give a geometrical explanation of this identity.

◊ 5.6.6. (a) Explain in detail why the  $n^{\text{th}}$  roots of 1 lie on the vertices of a regular  $n$ -gon. What is the angle between two consecutive sides?

(b) Explain why this is also true for the  $n^{\text{th}}$  roots of every non-zero complex number  $z \neq 0$ . Sketch a picture of the hexagon corresponding to  $\sqrt[6]{z}$  for a given  $z \neq 0$ .

◊ 5.6.7. In general, an  $n^{\text{th}}$  root of unity  $\zeta$  is called *primitive* if all the  $n^{\text{th}}$  roots of unity are obtained by raising it to successive powers:  $1, \zeta, \zeta^2, \zeta^3, \dots$ . (a) Find all primitive (i) fourth, (ii) fifth, (iii) ninth roots of unity. (b) Can you characterize all the primitive  $n^{\text{th}}$  roots of unity?

5.6.8. (a) In Example 5.27, the  $n = 4$  discrete Fourier coefficients of the function  $f(x) = 2\pi x - x^2$  were found to be real. Is this true when  $n = 16$ ? For general  $n$ ? (b) What property of a function  $f(x)$  will guarantee that its Fourier coefficients are real?

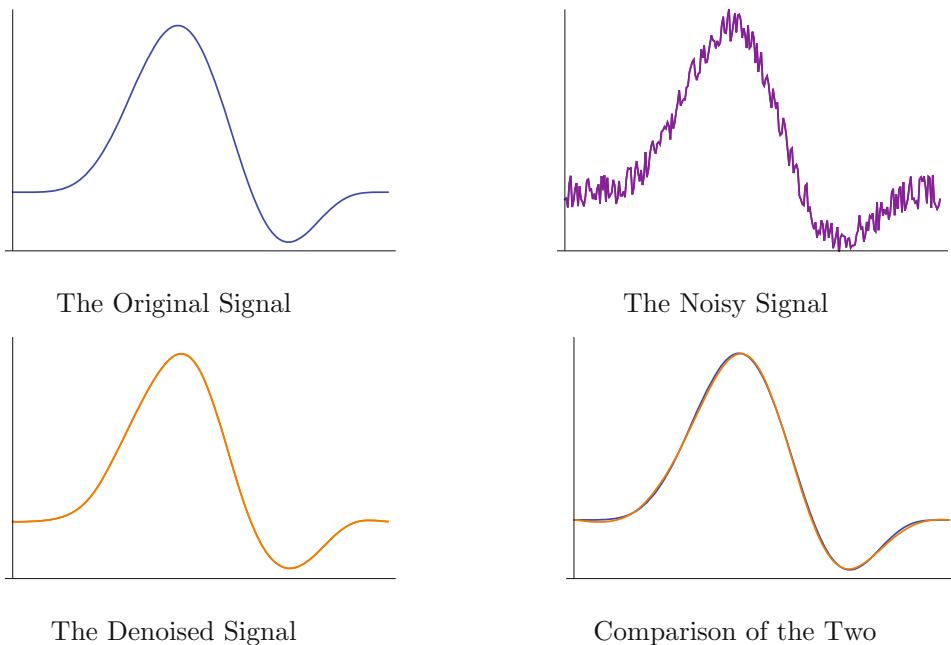


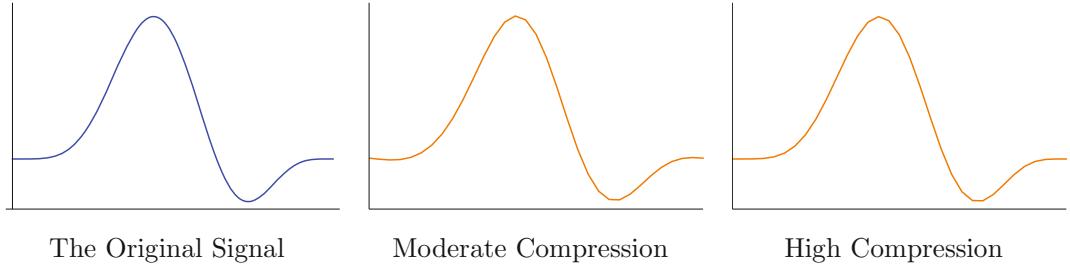
Figure 5.20. Denoising a Signal.

- ♡ 5.6.9. Let  $\mathbf{c} = (c_0, c_1, \dots, c_{n-1})^T \in \mathbb{C}^n$  be the vector of discrete Fourier coefficients corresponding to the sample vector  $\mathbf{f} = (f_0, f_1, \dots, f_{n-1})^T$ . (a) Explain why the sampled signal  $\mathbf{f} = F_n \mathbf{c}$  can be reconstructed by multiplying its Fourier coefficient vector by an  $n \times n$  matrix  $F_n$ . Write down  $F_2, F_3, F_4$ , and  $F_8$ . What is the general formula for the entries of  $F_n$ ? (b) Prove that, in general,  $F_n^{-1} = \frac{1}{n} F_n^\dagger = \frac{1}{n} \overline{F_n}^T$ , where  $\dagger$  denotes the Hermitian transpose defined in Exercise 4.3.25. (c) Prove that  $U_n = \frac{1}{\sqrt{n}} F_n$  is a unitary matrix, i.e.,  $U_n^{-1} = U_n^\dagger$ .

### Compression and Denoising

In a typical experimental signal, noise primarily affects the high-frequency modes, while the authentic features tend to appear in the low frequencies. Think of the hiss and static you hear on an AM radio station or a low-quality audio recording. Thus, a very simple, but effective, method for denoising a corrupted signal is to decompose it into its Fourier modes, as in (5.100), and then discard the high-frequency constituents. A similar idea underlies the Dolby recording system used on most movie soundtracks: during the recording process, the high-frequency modes are artificially boosted, so that scaling them back when the movie is shown in the theater has the effect of eliminating much of the extraneous noise. The one design issue is the specification of a cut-off between low and high frequency, that is, between signal and noise. This choice will depend upon the properties of the measured signal, and is left to the discretion of the signal processor.

A correct implementation of the denoising procedure is facilitated by using the unaliased forms (5.115, 116) of the trigonometric interpolant, in which the low-frequency summands appear only when  $|k|$  is small. In this version, to eliminate high-frequency components,



**Figure 5.21.** Compressing a Signal.

we replace the full summation by

$$q_l(x) = \sum_{k=-l}^l c_k e^{ikx}, \quad (5.119)$$

where  $l < \frac{1}{2}(n+1)$  specifies the selected cut-off frequency between signal and noise. The  $2l+1 \ll n$  low-frequency Fourier modes retained in (5.119) will, in favorable situations, capture the essential features of the original signal while simultaneously eliminating the high-frequency noise.

In Figure 5.20 we display a sample signal followed by the same signal corrupted by adding in random noise. We use  $n = 2^9 = 512$  sample points in the discrete Fourier representation, and to remove the noise, we retain only the  $2l+1 = 11$  lowest-frequency modes. In other words, instead of all  $n = 512$  Fourier coefficients  $c_{-256}, \dots, c_{-1}, c_0, c_1, \dots, c_{255}$ , we compute only the 11 lowest-order ones  $c_{-5}, \dots, c_5$ . Orthogonality is the key that allows us to do this! Summing up just those 11 exponentials produces the denoised signal  $q(x) = c_{-5} e^{-5ix} + \dots + c_5 e^{5ix}$ . To compare, we plot both the original signal and the denoised version on the same graph. In this case, the maximal deviation is less than .15 over the entire interval  $[0, 2\pi]$ .

The same idea underlies many data compression algorithms for audio recordings, digital images, and, particularly, video. The goal is efficient storage and/or transmission of the signal. As before, we expect all the important features to be contained in the low-frequency constituents, and so discarding the high-frequency terms will, in favorable situations, not lead to any noticeable degradation of the signal or image. Thus, to compress a signal (and, simultaneously, remove high-frequency noise), we retain only its low-frequency discrete Fourier coefficients. The signal is reconstructed by summing the associated discrete Fourier representation (5.119). A mathematical justification of Fourier-based compression algorithms relies on the fact that the Fourier coefficients of smooth functions tend rapidly to zero — the smoother the function, the faster the decay rate; see [61] for details. Thus, the small high-frequency Fourier coefficients will be of negligible importance.

In Figure 5.21, the same signal is compressed by retaining, respectively,  $2l+1 = 21$  and  $2l+1 = 7$  Fourier coefficients only instead of all  $n = 512$  that would be required for complete accuracy. For the case of moderate compression, the maximal deviation between the signal and the compressed version is less than  $1.5 \times 10^{-4}$  over the entire interval, while even the highly compressed version deviates by at most .05 from the original signal. Of course, the lack of any fine-scale features in this particular signal means that a very high compression can be achieved — the more complicated or detailed the original signal, the more Fourier modes need to be retained for accurate reproduction.

## Exercises

- ♠ 5.6.10. Construct the discrete Fourier coefficients for  $f(x) = \begin{cases} -x, & 0 \leq x \leq \frac{1}{3}\pi, \\ x - \frac{2}{3}\pi, & \frac{1}{3}\pi \leq x \leq \frac{4}{3}\pi, \\ -x + 2\pi, & \frac{4}{3}\pi \leq x \leq 2\pi. \end{cases}$

based on  $n = 128$  sample points. Then graph the reconstructed function when using the data compression algorithm that retains only the 11 and 21 lowest-frequency modes. Discuss what you observe.

- ♠ 5.6.11. Answer Exercise 5.6.10 when  $f(x) = (a) x; (b) x^2(2\pi - x)^2; (c) \begin{cases} \sin x, & 0 \leq x \leq \pi, \\ 0, & \pi \leq x \leq 2\pi. \end{cases}$
- ♣ 5.6.12. Let  $q_l(x)$  denote the trigonometric polynomial (5.119) obtained by summing the first  $2l + 1$  discrete Fourier modes. Suppose the criterion for compression of a signal  $f(x)$  is that  $\|f - q_l\|_\infty = \max\{|f(x) - q_l(x)| \mid 0 \leq x \leq 2\pi\} < \varepsilon$ . For the particular function in Exercise 5.6.10, how large do you need to choose  $k$  when  $\varepsilon = .1$ ?  $\varepsilon = .01$ ?  $\varepsilon = .001$ ?
- ♣ 5.6.13. Let  $f(x) = x(2\pi - x)$  be sampled on  $n = 128$  equally spaced points between 0 and  $2\pi$ . Use a random number generator with  $-1 \leq r_j \leq 1$  to add noise by replacing each sample value  $f_j = f(x_j)$  by  $g_j = f_j + \varepsilon r_j$ . Investigate, for different values of  $\varepsilon$ , how many discrete Fourier modes are required to reconstruct a reasonable denoised approximation to the original signal.

- ♠ 5.6.14. The signal in Figure 5.20 was obtained from the explicit formula

$$f(x) = -\frac{1}{5} \left( \frac{x(2\pi - x)}{10} \right)^5 (x + 1.5)(x + 2.5)(x - 4) + 1.7.$$

Noise was added by using a random number generator. Experiment with different intensities of noise and different numbers of sample points and discuss what you observe.

- ♣ 5.6.15. If we use the original form (5.100) of the discrete Fourier representation, we might be tempted to denoise/compress the signal by retaining only the first  $0 \leq k \leq l$  terms in the sum. Test this method on the signal in Exercise 5.6.10 and discuss what you observe.
- 5.6.16. *True or false:* If  $f(x)$  is real, the compressed/denoised signal (5.119) is a real trigonometric polynomial.

## The Fast Fourier Transform

While one may admire an algorithm for its intrinsic beauty, in the real world, the bottom line is always efficiency of implementation: the less total computation, the faster the processing, and hence the more extensive the range of applications. Orthogonality is the first and most important feature of many practical linear algebra algorithms, and is the critical feature of Fourier analysis. Still, even the power of orthogonality reaches its limits when it comes to dealing with truly large-scale problems such as three-dimensional medical imaging or video processing. In the early 1960's, James Cooley and John Tukey, [15], discovered<sup>†</sup> a much more efficient approach to the Discrete Fourier Transform, exploiting the rather special structure of the sampled exponential vectors. The resulting algorithm is

<sup>†</sup> In fact, the key ideas can be found in Gauss's hand computations in the early 1800's, but his insight was not fully appreciated until modern computers arrived on the scene.

known as the *Fast Fourier Transform*, often abbreviated FFT, and its discovery launched the modern revolution in digital signal and data processing, [9, 10].

In general, computing all the discrete Fourier coefficients (5.111) of an  $n$  times sampled signal requires a total of  $n^2$  complex multiplications and  $n^2 - n$  complex additions. Note also that each complex addition

$$z + w = (x + iy) + (u + iv) = (x + u) + i(y + v) \quad (5.120)$$

generally requires two real additions, while each complex multiplication

$$zw = (x + iy)(u + iv) = (xu - yv) + i(xv + yu) \quad (5.121)$$

requires 4 real multiplications and 2 real additions, or, by employing the alternative formula

$$xv + yu = (x + y)(u + v) - xu - yv \quad (5.122)$$

for the imaginary part, 3 real multiplications and 5 real additions. (The choice of formula (5.121) or (5.122) will depend upon the processor's relative speeds of multiplication and addition.) Similarly, given the Fourier coefficients  $c_0, \dots, c_{n-1}$ , reconstruction of the sampled signal via (5.100) requires  $n^2 - n$  complex multiplications and  $n^2 - n$  complex additions. As a result, both computations become quite labor-intensive for large  $n$ . Extending these ideas to multi-dimensional data only exacerbates the problem. The Fast Fourier Transform provides a shortcut around this computational bottleneck and thereby significantly extends the range of discrete Fourier analysis.

In order to explain the method without undue complication, we return to the original, aliased form of the discrete Fourier representation (5.100). (Once one understands how the FFT works, one can easily adapt the algorithm to the low-frequency version (5.116).) The seminal observation is that if the number of sample points  $n = 2m$  is even, then the primitive  $m^{\text{th}}$  root of unity  $\zeta_m = \sqrt[m]{1}$  equals the square of the primitive  $n^{\text{th}}$  root:  $\zeta_m = \zeta_n^2$ . We use this fact to split the summation (5.111) for the order  $n$  discrete Fourier coefficients into two parts, collecting the even and the odd powers of  $\zeta_n^k$ :

$$\begin{aligned} c_k &= \frac{1}{n} (f_0 + f_1 \zeta_n^{-k} + f_2 \zeta_n^{-2k} + \dots + f_{n-1} \zeta_n^{-(n-1)k}) \\ &= \frac{1}{n} (f_0 + f_2 \zeta_n^{-2k} + f_4 \zeta_n^{-4k} + \dots + f_{2m-2} \zeta_n^{-(2m-2)k}) + \\ &\quad + \zeta_n^{-k} \frac{1}{n} (f_1 + f_3 \zeta_n^{-2k} + f_5 \zeta_n^{-4k} + \dots + f_{2m-1} \zeta_n^{-(2m-2)k}) \\ &= \frac{1}{2} \left\{ \frac{1}{m} (f_0 + f_2 \zeta_m^{-k} + f_4 \zeta_m^{-2k} + \dots + f_{2m-2} \zeta_m^{-(m-1)k}) \right\} + \\ &\quad + \frac{\zeta_n^{-k}}{2} \left\{ \frac{1}{m} (f_1 + f_3 \zeta_m^{-k} + f_5 \zeta_m^{-2k} + \dots + f_{2m-1} \zeta_m^{-(m-1)k}) \right\}. \end{aligned} \quad (5.123)$$

Now, observe that the expressions in braces are the order  $m$  Fourier coefficients for the sample data

$$\begin{aligned} \mathbf{f}^e &= (f_0, f_2, f_4, \dots, f_{2m-2})^T = (f(x_0), f(x_2), f(x_4), \dots, f(x_{2m-2}))^T, \\ \mathbf{f}^o &= (f_1, f_3, f_5, \dots, f_{2m-1})^T = (f(x_1), f(x_3), f(x_5), \dots, f(x_{2m-1}))^T. \end{aligned} \quad (5.124)$$

Note that  $\mathbf{f}^e$  is obtained by sampling  $f(x)$  on the *even* sample points  $x_{2j}$ , while  $\mathbf{f}^o$  is obtained by sampling the same function  $f(x)$ , but now at the *odd* sample points  $x_{2j+1}$ . In

other words, we are splitting the original sampled signal into two “half-sampled” signals obtained by sampling on every other point. The even and odd Fourier coefficients are

$$\begin{aligned} c_k^e &= \frac{1}{m} (f_0 + f_2 \zeta_m^{-k} + f_4 \zeta_m^{-2k} + \cdots + f_{2m-2} \zeta_m^{-(m-1)k}), & k = 0, \dots, m-1. \\ c_k^o &= \frac{1}{m} (f_1 + f_3 \zeta_m^{-k} + f_5 \zeta_m^{-2k} + \cdots + f_{2m-1} \zeta_m^{-(m-1)k}), \end{aligned} \quad (5.125)$$

Since they contain just  $m$  data values, both the even and odd samples require only  $m$  distinct Fourier coefficients, and we adopt the identification

$$c_{k+m}^e = c_k^e, \quad c_{k+m}^o = c_k^o, \quad k = 0, \dots, m-1. \quad (5.126)$$

Therefore, the order  $n = 2m$  discrete Fourier coefficients (5.123) can be constructed from a pair of order  $m$  discrete Fourier coefficients via

$$c_k = \frac{1}{2} (c_k^e + \zeta_n^{-k} c_k^o), \quad k = 0, \dots, n-1. \quad (5.127)$$

Now if  $m = 2l$  is also even, then we can play the same game on the order  $m$  Fourier coefficients (5.125), reconstructing each of them from a pair of order  $l$  discrete Fourier coefficients — obtained by sampling the signal at every fourth point. If  $n = 2^r$  is a power of 2, then this game can be played all the way back to the start, beginning with the trivial order 1 discrete Fourier representation, which just samples the function at a single point. The result is the desired algorithm. After some rearrangement of the basic steps, we arrive at the Fast Fourier Transform, which we now present in its final form.

We begin with a sampled signal on  $n = 2^r$  sample points. To efficiently program the Fast Fourier Transform, it helps to write out each index  $0 \leq j < 2^r$  in its binary (as opposed to decimal) representation

$$j = j_{r-1} j_{r-2} \cdots j_2 j_1 j_0, \quad \text{where} \quad j_\nu = 0 \text{ or } 1; \quad (5.128)$$

the notation is shorthand for its  $r$  digit binary expansion

$$j = j_0 + 2j_1 + 4j_2 + 8j_3 + \cdots + 2^{r-1} j_{r-1}.$$

We then define the *bit reversal* map

$$\rho(j_{r-1} j_{r-2} \cdots j_2 j_1 j_0) = j_0 j_1 j_2 \cdots j_{r-2} j_{r-1}. \quad (5.129)$$

For instance, if  $r = 5$ , and  $j = 13$ , with 5 digit binary representation 01101, then  $\rho(j) = 22$  has the reversed binary representation 10110. Note especially that the bit reversal map  $\rho = \rho_r$  depends upon the original choice of  $r = \log_2 n$ .

Secondly, for each  $0 \leq k < r$ , define the maps

$$\begin{aligned} \alpha_k(j) &= j_{r-1} \cdots j_{k+1} 0 j_{k-1} \cdots j_0, \\ \beta_k(j) &= j_{r-1} \cdots j_{k+1} 1 j_{k-1} \cdots j_0 = \alpha_k(j) + 2^k, \end{aligned} \quad \text{for } j = j_{r-1} j_{r-2} \cdots j_1 j_0. \quad (5.130)$$

In other words,  $\alpha_k(j)$  sets the  $k^{\text{th}}$  binary digit of  $j$  to 0, while  $\beta_k(j)$  sets it to 1. In the preceding example,  $\alpha_2(13) = 9$ , with binary form 01001, while  $\beta_2(13) = 13$  with binary form 01101. The bit operations (5.129, 130) are especially easy to implement on modern binary computers.

Given a sampled signal  $f_0, \dots, f_{n-1}$ , its discrete Fourier coefficients  $c_0, \dots, c_{n-1}$  are computed by the following iterative algorithm:

$$\begin{aligned} c_j^{(0)} &= f_{\rho(j)}, & c_j^{(k+1)} &= \frac{1}{2} (c_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^{-j} c_{\beta_k(j)}^{(k)}), & j &= 0, \dots, n-1, \\ & & & & & k = 0, \dots, r-1, \end{aligned} \quad (5.131)$$

in which  $\zeta_{2^{k+1}}$  is the primitive  $2^{k+1}$  root of unity. The final output of the iterative procedure, namely

$$c_j = c_j^{(r)}, \quad j = 0, \dots, n-1, \quad (5.132)$$

are the discrete Fourier coefficients of our signal. The preprocessing step of the algorithm, where we define  $c_j^{(0)}$ , produces a more convenient rearrangement of the sample values. The subsequent steps successively combine the Fourier coefficients of the appropriate even and odd sampled subsignals, reproducing (5.123) in a different notation. The following example should help make the overall process clearer.

**Example 5.28.** Consider the case  $r = 3$ , and so our signal has  $n = 2^3 = 8$  sampled values  $f_0, f_1, \dots, f_7$ . We begin the process by rearranging the sample values

$$c_0^{(0)} = f_0, \quad c_1^{(0)} = f_4, \quad c_2^{(0)} = f_2, \quad c_3^{(0)} = f_6, \quad c_4^{(0)} = f_1, \quad c_5^{(0)} = f_5, \quad c_6^{(0)} = f_3, \quad c_7^{(0)} = f_7,$$

in the order specified by the bit reversal map  $\rho$ . For instance,  $\rho(3) = 6$ , or, in binary notation,  $\rho(011) = 110$ .

The first stage of the iteration is based on  $\zeta_2 = -1$ . Equation (5.131) gives

$$\begin{aligned} c_0^{(1)} &= \frac{1}{2}(c_0^{(0)} + c_1^{(0)}), & c_1^{(1)} &= \frac{1}{2}(c_0^{(0)} - c_1^{(0)}), & c_2^{(1)} &= \frac{1}{2}(c_2^{(0)} + c_3^{(0)}), & c_3^{(1)} &= \frac{1}{2}(c_2^{(0)} - c_3^{(0)}), \\ c_4^{(1)} &= \frac{1}{2}(c_4^{(0)} + c_5^{(0)}), & c_5^{(1)} &= \frac{1}{2}(c_4^{(0)} - c_5^{(0)}), & c_6^{(1)} &= \frac{1}{2}(c_6^{(0)} + c_7^{(0)}), & c_7^{(1)} &= \frac{1}{2}(c_6^{(0)} - c_7^{(0)}), \end{aligned}$$

where we combine successive pairs of the rearranged sample values. The second stage of the iteration has  $k = 1$  with  $\zeta_4 = i$ . We obtain

$$\begin{aligned} c_0^{(2)} &= \frac{1}{2}(c_0^{(1)} + c_2^{(1)}), & c_1^{(2)} &= \frac{1}{2}(c_1^{(1)} - i c_3^{(1)}), & c_2^{(2)} &= \frac{1}{2}(c_0^{(1)} - c_2^{(1)}), & c_3^{(2)} &= \frac{1}{2}(c_1^{(1)} + i c_3^{(1)}), \\ c_4^{(2)} &= \frac{1}{2}(c_4^{(1)} + c_6^{(1)}), & c_5^{(2)} &= \frac{1}{2}(c_5^{(1)} - i c_7^{(1)}), & c_6^{(2)} &= \frac{1}{2}(c_4^{(1)} - c_6^{(1)}), & c_7^{(2)} &= \frac{1}{2}(c_5^{(1)} + i c_7^{(1)}). \end{aligned}$$

Note that the indices of the combined pairs of coefficients differ by 2. In the last step, where  $k = 2$  and  $\zeta_8 = \frac{\sqrt{2}}{2}(1 + i)$ , we combine coefficients whose indices differ by  $4 = 2^2$ ; the final output

$$\begin{aligned} c_0 &= c_0^{(3)} = \frac{1}{2}(c_0^{(2)} + c_4^{(2)}), & c_4 &= c_4^{(3)} = \frac{1}{2}(c_0^{(2)} - c_4^{(2)}), \\ c_1 &= c_1^{(3)} = \frac{1}{2} \left( c_1^{(2)} + \frac{\sqrt{2}}{2}(1 - i)c_5^{(2)} \right), & c_5 &= c_5^{(3)} = \frac{1}{2} \left( c_1^{(2)} - \frac{\sqrt{2}}{2}(1 - i)c_5^{(2)} \right), \\ c_2 &= c_2^{(3)} = \frac{1}{2} \left( c_2^{(2)} - i c_6^{(2)} \right), & c_6 &= c_6^{(3)} = \frac{1}{2} \left( c_2^{(2)} + i c_6^{(2)} \right), \\ c_3 &= c_3^{(3)} = \frac{1}{2} \left( c_3^{(2)} - \frac{\sqrt{2}}{2}(1 + i)c_7^{(2)} \right), & c_7 &= c_7^{(3)} = \frac{1}{2} \left( c_3^{(2)} + \frac{\sqrt{2}}{2}(1 + i)c_7^{(2)} \right), \end{aligned}$$

is the complete set of discrete Fourier coefficients.

Let us count the number of arithmetic operations required in the Fast Fourier Transform algorithm. At each stage in the computation, we must perform  $n = 2^r$  complex additions/subtractions and the same number of complex multiplications. (Actually, the number of multiplications is slightly smaller, since multiplications by  $\pm 1$  and  $\pm i$  are extremely simple. However, this does not significantly alter the final operations count.) There are  $r = \log_2 n$  stages, and so we require a total of  $r n = n \log_2 n$  complex additions/subtractions and the same number of multiplications. Now, when  $n$  is large,  $n \log_2 n$  is *significantly* smaller than  $n^2$ , which is the number of operations required for the direct algorithm. For instance, if  $n = 2^{10} = 1,024$ , then  $n^2 = 1,048,576$ , while  $n \log_2 n = 10,240$  — a net savings of 99%. As a result, many large-scale computations that would be intractable using the direct approach are immediately brought into the realm of feasibility. This is the reason why all modern

implementations of the Discrete Fourier Transform are based on the FFT algorithm and its variants.

The reconstruction of the signal from the discrete Fourier coefficients  $c_0, \dots, c_{n-1}$  is speeded up in exactly the same manner. The only differences are that we replace  $\zeta_n^{-1} = \overline{\zeta_n}$  by  $\zeta_n$ , and drop the factors of  $\frac{1}{2}$ , since there is no need to divide by  $n$  in the final result (5.100). Therefore, we apply the slightly modified iterative procedure

$$f_j^{(0)} = c_{\rho(j)}, \quad f_j^{(k+1)} = f_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^j f_{\beta_k(j)}^{(k)}, \quad \begin{aligned} j &= 0, \dots, n-1, \\ k &= 0, \dots, r-1, \end{aligned} \quad (5.133)$$

and finish with

$$f(x_j) = f_j = f_j^{(r)}, \quad j = 0, \dots, n-1. \quad (5.134)$$

**Example 5.29.** The reconstruction formulas in the case of  $n = 8 = 2^3$  Fourier coefficients  $c_0, \dots, c_7$ , which were computed in Example 5.28, can be implemented as follows. First, we rearrange the Fourier coefficients in bit reversed order:

$$f_0^{(0)} = c_0, \quad f_1^{(0)} = c_4, \quad f_2^{(0)} = c_2, \quad f_3^{(0)} = c_6, \quad f_4^{(0)} = c_1, \quad f_5^{(0)} = c_5, \quad f_6^{(0)} = c_3, \quad f_7^{(0)} = c_7.$$

Then we begin combining them in successive pairs:

$$\begin{aligned} f_0^{(1)} &= f_0^{(0)} + f_1^{(0)}, & f_1^{(1)} &= f_0^{(0)} - f_1^{(0)}, & f_2^{(1)} &= f_2^{(0)} + f_3^{(0)}, & f_3^{(1)} &= f_2^{(0)} - f_3^{(0)}, \\ f_4^{(1)} &= f_4^{(0)} + f_5^{(0)}, & f_5^{(1)} &= f_4^{(0)} - f_5^{(0)}, & f_6^{(1)} &= f_6^{(0)} + f_7^{(0)}, & f_7^{(1)} &= f_6^{(0)} - f_7^{(0)}. \end{aligned}$$

Next,

$$\begin{aligned} f_0^{(2)} &= f_0^{(1)} + f_2^{(1)}, & f_1^{(2)} &= f_1^{(1)} + i f_3^{(1)}, & f_2^{(2)} &= f_0^{(1)} - f_2^{(1)}, & f_3^{(2)} &= f_1^{(1)} - i f_3^{(1)}, \\ f_4^{(2)} &= f_4^{(1)} + f_6^{(1)}, & f_5^{(2)} &= f_5^{(1)} + i f_7^{(1)}, & f_6^{(2)} &= f_4^{(1)} - f_6^{(1)}, & f_7^{(2)} &= f_5^{(1)} - i f_7^{(1)}. \end{aligned}$$

Finally, the sampled signal values are

$$\begin{aligned} f(x_0) &= f_0^{(3)} = f_0^{(2)} + f_4^{(2)}, & f(x_4) &= f_4^{(3)} = f_0^{(2)} - f_4^{(2)}, \\ f(x_1) &= f_1^{(3)} = f_1^{(2)} + \frac{\sqrt{2}}{2} (1 + i) f_5^{(2)}, & f(x_5) &= f_5^{(3)} = f_1^{(2)} - \frac{\sqrt{2}}{2} (1 + i) f_5^{(2)}, \\ f(x_2) &= f_2^{(3)} = f_2^{(2)} + i f_6^{(2)}, & f(x_6) &= f_6^{(3)} = f_2^{(2)} - i f_6^{(2)}, \\ f(x_3) &= f_3^{(3)} = f_3^{(2)} - \frac{\sqrt{2}}{2} (1 - i) f_7^{(2)}, & f(x_7) &= f_7^{(3)} = f_3^{(2)} + \frac{\sqrt{2}}{2} (1 - i) f_7^{(2)}. \end{aligned}$$

## Exercises

♣ 5.6.17. Use the Fast Fourier Transform to find the discrete Fourier coefficients for the the following functions using the indicated number of sample points. Carefully indicate each step in your analysis.

$$(a) \frac{x}{\pi}, \quad n = 4; \quad (b) \sin x, \quad n = 8; \quad (c) |x - \pi|, \quad n = 8; \quad (d) \operatorname{sign}(x - \pi), \quad n = 16.$$

♣ 5.6.18. Use the Inverse Fast Fourier Transform to reassemble the sampled function data corresponding to the following discrete Fourier coefficients. Carefully indicate each step in your analysis.

$$(a) c_0 = c_2 = 1, c_1 = c_3 = -1, \quad (b) c_0 = c_1 = c_4 = 2, c_2 = c_6 = 0, c_3 = c_5 = c_7 = -1.$$

- ♡ 5.6.19. In this exercise, we show how the Fast Fourier Transform is equivalent to a certain matrix factorization. Let  $\mathbf{c} = (c_0, c_1, \dots, c_7)^T$  be the vector of Fourier coefficients, and let  $\mathbf{f}^{(k)} = (f_0^{(k)}, f_1^{(k)}, \dots, f_7^{(k)})^T$  for  $k = 0, 1, 2, 3$ , be the vectors containing the coefficients defined in the reconstruction algorithm Example 5.29. (a) Show that  $\mathbf{f}^{(0)} = M_0\mathbf{c}$ ,  $\mathbf{f}^{(1)} = M_1\mathbf{f}^{(0)}$ ,  $\mathbf{f}^{(2)} = M_2\mathbf{f}^{(1)}$ ,  $\mathbf{f} = \mathbf{f}^{(3)} = M_3\mathbf{f}^{(2)}$ , where  $M_0, M_1, M_2, M_3$  are  $8 \times 8$  matrices. Write down their explicit forms. (b) Explain why the matrix product  $F_8 = M_3 M_2 M_1 M_0$  reproduces the Fourier matrix derived in Exercise 5.6.9. Check the factorization directly. (c) Write down the corresponding matrix factorization for the direct algorithm of Example 5.28.
-



# Chapter 6

# Equilibrium

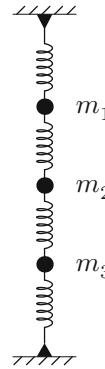
In this chapter, we will apply what we have learned so far to the analysis of equilibrium configurations and stability of mechanical structures and electrical networks. Both physical problems fit into a common, and surprisingly general, mathematical framework. The physical laws of equilibrium mechanics and circuits lead to linear algebraic systems whose coefficient matrix is of positive (semi-)definite Gram form. The positive definite cases correspond to stable structures and networks, which can support any applied forcing or external current, producing a unique, stable equilibrium solution that can be characterized by an energy minimization principle. On the other hand, systems with semi-definite coefficient matrices model unstable structures and networks that are unable to remain in equilibrium except under very special configurations of external forces. In the case of mechanical structures, the instabilities are of two types: rigid motions, in which the structure moves while maintaining its overall geometrical shape, and mechanisms, in which it spontaneously deforms in the absence of any applied force. The same linear algebra framework, but now reformulated for infinite-dimensional function space, also characterizes the boundary value problems for both ordinary and partial differential equation that model the equilibria of continuous media, including bars, beams, solid bodies, and many other systems arising throughout physics and engineering, [61, 79].

The starting point is a linear chain of masses interconnected by springs and constrained to move only in the longitudinal direction. Our general mathematical framework is already manifest in this rather simple mechanical system. In the second section, we discuss simple electrical networks consisting of resistors, current sources and/or batteries, interconnected by a network of wires. Here, the resulting Gram matrix is known as the graph Laplacian, which plays an increasingly important role in modern data analysis and network theory. Finally, we treat small (so as to remain in a linear modeling regime) displacements of two- and three-dimensional structures constructed out of elastic bars. In all cases, we consider only the equilibrium solutions. Dynamical (time-varying) processes for each of these physical systems are governed by linear systems of ordinary differential equations, to be formulated and analyzed in Chapter 10.

## 6.1 Springs and Masses

A *mass–spring chain* consists of  $n$  masses  $m_1, m_2, \dots, m_n$  arranged in a straight line. Each mass is connected to its immediate neighbor(s) by springs. Moreover, the chain may be connected at one or both ends to a fixed support by a spring — or may even be completely free, e.g., floating in outer space. For specificity, let us first look at the case when both ends of the chain are attached to unmoving supports, as illustrated in Figure 6.1

We assume that the masses are arranged in a vertical line, and order them from top to bottom. For simplicity, we will only allow the masses to move in the vertical direction, that is, we restrict to a one-dimensional motion. (Section 6.3 deals with the more complicated two- and three-dimensional situations.)



**Figure 6.1.** A Mass-Spring Chain with Fixed Ends.

If we subject some or all of the masses to an external force, e.g., gravity, then the system will move<sup>†</sup> to a new equilibrium position. The resulting position of the  $i^{\text{th}}$  mass is measured by its *displacement*  $u_i$  from its original position, which, since we are only allowing vertical motion, is a scalar quantity. Referring to Figure 6.1, we use the convention that  $u_i > 0$  if the mass has moved downwards, and  $u_i < 0$  if it has moved upwards. Our goal is to determine the new equilibrium configuration of the chain under the prescribed forcing, that is, to set up and solve a system of equations for the displacements  $u_1, \dots, u_n$ .

As sketched in Figure 6.2, let  $e_j$  denote the *elongation* of the  $j^{\text{th}}$  spring, which connects mass  $m_{j-1}$  to mass  $m_j$ . By “elongation”, we mean how far the spring has been stretched, so that  $e_j > 0$  if the spring is longer than its reference length, while  $e_j < 0$  if the spring has been compressed. The elongations of the internal springs can be determined directly from the displacements of the masses at each end according to the geometric formula

$$e_j = u_j - u_{j-1}, \quad j = 2, \dots, n, \quad (6.1)$$

while, for the top and bottom springs,

$$e_1 = u_1, \quad e_{n+1} = -u_n, \quad (6.2)$$

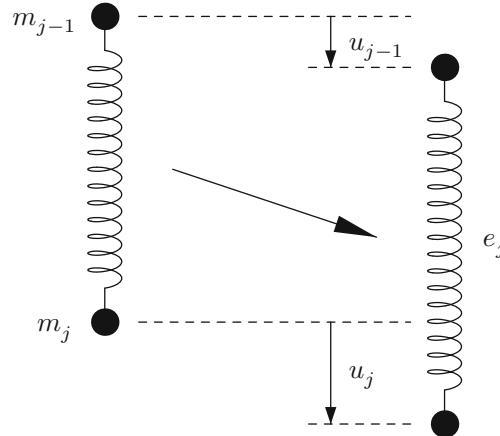
since the supports are not allowed to move. We write the elongation equations (6.1–2) in matrix form

$$\mathbf{e} = A \mathbf{u}, \quad (6.3)$$

where  $\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n+1} \end{pmatrix}$  is the *elongation vector*,  $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$  is the *displacement vector*, and

---

<sup>†</sup> The differential equations governing its dynamical behavior during the motion will be the subject of Chapter 10. Damping or frictional effects will cause the system to eventually settle down into a stable equilibrium configuration, if such exists.



**Figure 6.2.** Elongation of a Spring.

the coefficient matrix

$$A = \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \\ & & & & & -1 \end{pmatrix} \quad (6.4)$$

has size  $(n + 1) \times n$ , with only its non-zero entries being indicated. We refer to  $A$  as the *reduced incidence matrix*<sup>†</sup> for the mass–spring chain. The incidence matrix effectively encodes the underlying geometry of the system, including the fixed “boundary conditions” at the top and the bottom.

The next step is to relate the elongation  $e_j$  experienced by the  $j^{\text{th}}$  spring to its internal force  $y_j$ . This is the basic *constitutive assumption*, which relates geometry to kinematics. In the present case, we suppose that the springs are not stretched (or compressed) particularly far. Under this assumption, *Hooke’s Law*, named in honor of the seventeenth-century English scientist and inventor Robert Hooke, states that the internal force is directly proportional to the elongation — the more you stretch a spring, the more it tries to pull you back. Thus,

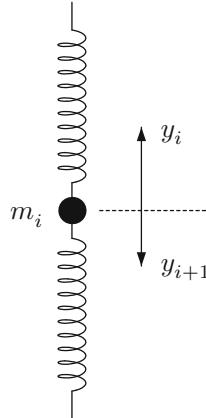
$$y_j = c_j e_j, \quad (6.5)$$

where the constant of proportionality  $c_j > 0$  measures the spring’s *stiffness*. Hard springs have large stiffness and so takes a large force to stretch, whereas soft springs have a small, but still positive, stiffness. We will also write the constitutive equations (6.5) in matrix form

$$\mathbf{y} = C \mathbf{e}, \quad (6.6)$$

---

<sup>†</sup> The connection with the incidence matrix of a graph, as introduced in Section 2.6, will become evident in the following Section 6.2.



**Figure 6.3.** Force Balance.

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n+1} \end{pmatrix}, \quad C = \begin{pmatrix} c_1 & & & \\ & c_2 & & \\ & & \ddots & \\ & & & c_{n+1} \end{pmatrix}$$

are the internal force vector and the matrix of spring stiffnesses. Note particularly that  $C$  is a diagonal matrix, and, more importantly, positive definite,  $C > 0$ , since all its diagonal entries are strictly positive.

Finally, the forces must balance if the system is to remain in equilibrium. In this simplified model, the external forces act only on the masses, and not on the springs. Let  $f_i$  denote the external force on the  $i^{\text{th}}$  mass  $m_i$ . We also measure force in the downward direction, so  $f_i > 0$  means that the force is pulling the  $i^{\text{th}}$  mass downward. (In particular, gravity would induce a positive force on each mass.) If the  $i^{\text{th}}$  spring is stretched, it will exert an upward force on  $m_i$ , while if the  $(i+1)^{\text{st}}$  spring is stretched, it will pull  $m_i$  downward. Therefore, the balance of forces on  $m_i$  requires that

$$f_i = y_i - y_{i+1}. \quad (6.7)$$

The vectorial form of the force balance law is

$$\mathbf{f} = A^T \mathbf{y}, \quad (6.8)$$

where  $\mathbf{f} = (f_1, \dots, f_n)^T$ . The remarkable fact is that the force balance coefficient matrix

$$A^T = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & 1 & -1 & \\ & & & & \ddots & \ddots \\ & & & & & 1 & -1 \end{pmatrix} \quad (6.9)$$

is the *transpose* of the reduced incidence matrix (6.4) for the chain. This connection between geometry and force balance turns out to be of almost universal applicability, and

is the reason underlying the positivity of the final coefficient matrix in the resulting system of equilibrium equations.

Summarizing, the basic geometrical and physical properties of our mechanical system lead us to the full system of equilibrium equations (6.3, 6, 8) relating its displacements  $\mathbf{u}$ , elongations  $\mathbf{e}$ , internal forces  $\mathbf{y}$ , and external forces  $\mathbf{f}$ :

$$\mathbf{e} = A\mathbf{u}, \quad \mathbf{y} = C\mathbf{e}, \quad \mathbf{f} = A^T\mathbf{y}. \quad (6.10)$$

These equations imply  $\mathbf{f} = A^T\mathbf{y} = A^T C \mathbf{e} = A^T C A \mathbf{u}$ , and hence can be combined into a single linear system

$$K\mathbf{u} = \mathbf{f}, \quad \text{where} \quad K = A^T C A \quad (6.11)$$

is called the *stiffness matrix* associated with the entire mass–spring chain. In the particular case under consideration,

$$K = \begin{pmatrix} c_1 + c_2 & -c_2 & & & \\ -c_2 & c_2 + c_3 & -c_3 & & \\ & -c_3 & c_3 + c_4 & -c_4 & \\ & & -c_4 & c_4 + c_5 & -c_5 \\ & & & \ddots & \ddots & \ddots \\ & & & & -c_{n-1} & c_{n-1} + c_n & -c_n \\ & & & & & -c_n & c_n + c_{n+1} \end{pmatrix} \quad (6.12)$$

has a very simple symmetric, tridiagonal form. As such, we can use the tridiagonal solution algorithm of Section 1.7 to rapidly solve the linear system (6.11) for the displacements of the masses. Once we have solved (6.11) for the displacements  $\mathbf{u}$  we can then compute the resulting elongations  $\mathbf{e}$  and internal forces  $\mathbf{y}$  by substituting into the original system (6.10).

**Example 6.1.** Let us consider the particular case of  $n = 3$  masses connected by identical springs with unit spring constant. Thus,  $c_1 = c_2 = c_3 = c_4 = 1$ , and  $C = \text{diag}(1, 1, 1, 1) = I$  is the  $4 \times 4$  identity matrix. The  $3 \times 3$  stiffness matrix is then

$$K = A^T A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

A straightforward Gaussian Elimination produces the  $K = LDL^T$  factorization

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & \frac{4}{3} \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 \end{pmatrix}. \quad (6.13)$$

With this in hand, we can solve the basic equilibrium equations  $K\mathbf{u} = \mathbf{f}$  by the usual Forward and Back Substitution algorithm.

**Remark.** Even though we construct  $K = A^T C A$  and then factor it as  $K = LDL^T$ , there is no direct algorithm to get from  $A$  and  $C$  to  $L$  and  $D$ , which, typically, are matrices of different sizes.

Suppose, for example, we pull the middle mass downwards with a unit force, so  $f_2 = 1$  while  $f_1 = f_3 = 0$ . Then  $\mathbf{f} = (0, 1, 0)^T$ , and the solution to the equilibrium equations

(6.11) is  $\mathbf{u} = (\frac{1}{2}, 1, \frac{1}{2})^T$ , whose entries prescribe the mass displacements. Observe that all three masses have moved down, with the middle mass moving twice as far as the other two. The corresponding spring elongations and internal forces are obtained by matrix multiplication

$$\mathbf{y} = \mathbf{e} = A\mathbf{u} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix},$$

since  $C = \mathbf{I}$ . Thus, the top two springs are elongated, while the bottom two are compressed, all by an equal amount.

Similarly, if all the masses are equal,  $m_1 = m_2 = m_3 = m$ , then the solution under a constant downwards gravitational force  $\mathbf{f} = (mg, mg, mg)^T$  is

$$\mathbf{u} = K^{-1} \begin{pmatrix} mg \\ mg \\ mg \end{pmatrix} = \begin{pmatrix} \frac{3}{2}mg \\ 2mg \\ \frac{3}{2}mg \end{pmatrix}, \quad \text{and} \quad \mathbf{y} = \mathbf{e} = A\mathbf{u} = \begin{pmatrix} \frac{3}{2}mg \\ \frac{1}{2}mg \\ -\frac{1}{2}mg \\ -\frac{3}{2}mg \end{pmatrix}.$$

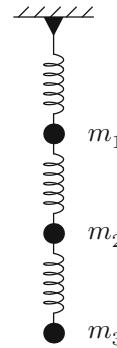
Now, the middle mass has only moved 33% farther than the others, whereas the top and bottom springs are experiencing three times as much elongation/compression as the middle two springs.

An important observation is that we *cannot* determine the internal forces  $\mathbf{y}$  or elongations  $\mathbf{e}$  directly from the force balance law (6.8), because the transposed matrix  $A^T$  is not square, and so the system  $\mathbf{f} = A^T\mathbf{y}$  does not have a unique solution. We must first compute the displacements  $\mathbf{u}$  by solving the full equilibrium equations (6.11), and then use the resulting displacements to reconstruct the elongations and internal forces. Such systems are referred to as *statically indeterminate*.

The behavior of the system will depend on both the forcing and the boundary conditions. Suppose, by way of contrast, that we fix only the top of the chain to a support, and leave the bottom mass hanging freely, as in [Figure 6.4](#). The geometric relation between the displacements and the elongations has the same form (6.3) as before, but the reduced incidence matrix is slightly altered:

$$A = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \end{pmatrix}. \quad (6.14)$$

This matrix has size  $n \times n$  and is obtained from the preceding example (6.4) by eliminating the last row corresponding to the missing bottom spring. The constitutive equations are still governed by Hooke's law  $\mathbf{y} = C\mathbf{e}$ , as in (6.6), with  $C = \text{diag}(c_1, \dots, c_n)$  the  $n \times n$  diagonal matrix of spring stiffnesses. Finally, the force balance equations are also found to have the same general form  $\mathbf{f} = A^T\mathbf{y}$  as in (6.8), but with the transpose of the revised incidence matrix (6.14). In conclusion, the equilibrium equations  $K\mathbf{u} = \mathbf{f}$  have an identical



**Figure 6.4.** A Mass–Spring Chain with One Free End.

form (6.11), based on the revised stiffness matrix

$$K = A^T C A = \begin{pmatrix} c_1 + c_2 & -c_2 & & & & \\ -c_2 & c_2 + c_3 & -c_3 & & & \\ & -c_3 & c_3 + c_4 & -c_4 & & \\ & & -c_4 & c_4 + c_5 & -c_5 & \\ & & & \ddots & \ddots & \ddots \\ & & & & -c_{n-1} & c_{n-1} + c_n & -c_n \\ & & & & & -c_n & c_n \end{pmatrix} \quad (6.15)$$

Only the bottom right entry differs from the fixed end matrix (6.12).

This system is called *statically determinate*, because the incidence matrix  $A$  is square and nonsingular, and so it is possible to solve the force balance law (6.8) directly for the internal forces  $\mathbf{y} = A^{-T}\mathbf{f}$  without having to solve the full equilibrium equations for the displacements  $\mathbf{u}$  before computing the internal forces  $\mathbf{y} = CA\mathbf{u}$ .

**Example 6.2.** For a three mass chain with one free end and equal unit spring constants  $c_1 = c_2 = c_3 = 1$ , the stiffness matrix is

$$K = A^T A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Pulling the middle mass downwards with a unit force, whereby  $\mathbf{f} = (0, 1, 0)^T$ , results in the displacements

$$\mathbf{u} = K^{-1}\mathbf{f} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \quad \text{so that} \quad \mathbf{y} = \mathbf{e} = A\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

In this configuration, the bottom two masses have moved by the same amount, which is twice as far as the top mass. Because we are pulling only on the middle mass, the bottom spring hangs free and experiences no elongation, whereas the top two springs are stretched by the same amount.

Similarly, for a chain of equal masses subject to a constant downwards gravitational

force  $\mathbf{f} = (mg, mg, mg)^T$ , the equilibrium position is

$$\mathbf{u} = K^{-1} \begin{pmatrix} mg \\ mg \\ mg \end{pmatrix} = \begin{pmatrix} 3mg \\ 5mg \\ 6mg \end{pmatrix}, \quad \text{and} \quad \mathbf{y} = \mathbf{e} = A\mathbf{u} = \begin{pmatrix} 3mg \\ 2mg \\ mg \end{pmatrix}.$$

Note how much farther the masses have moved now that the restraining influence of the bottom support has been removed. The top spring is experiencing the most elongation, and is thus the most likely to break, because it must support all three masses.

## Exercises

- 6.1.1. A mass–spring chain consists of two masses connected to two fixed supports. The spring constants are  $c_1 = c_3 = 1$  and  $c_2 = 2$ . (a) Find the stiffness matrix  $K$ . (b) Solve the equilibrium equations  $K\mathbf{u} = \mathbf{f}$  when  $\mathbf{f} = (4, 3)^T$ . (c) Which mass moved the farthest? (d) Which spring has been stretched the most? Compressed the most?
- 6.1.2. Solve Exercise 6.1.1 when the first and second springs are interchanged,  $c_1 = 2$ ,  $c_2 = c_3 = 1$ . Which of your conclusions changed?
- 6.1.3. Redo Exercises 6.1.1–2 when the bottom support and spring are removed.
- 6.1.4. A mass–spring chain consists of four masses suspended between two fixed supports. The spring stiffnesses are  $c_1 = 1$ ,  $c_2 = \frac{1}{2}$ ,  $c_3 = \frac{2}{3}$ ,  $c_4 = \frac{1}{2}$ ,  $c_5 = 1$ . (a) Determine the equilibrium positions of the masses and the elongations of the springs when the external force is  $\mathbf{f} = (0, 1, 1, 0)^T$ . Is your solution unique? (b) Suppose we fix only the top support. Solve the problem with the same data and compare your results.
- 6.1.5. (a) Show that, in a mass–spring chain with two fixed ends, under any external force, the average elongation of the springs is zero:  $\frac{1}{n+1}(e_1 + \dots + e_{n+1}) = 0$ . (b) What can you say about the average elongation of the springs in a chain with one fixed end?
- ◇ 6.1.6. Suppose we subject the  $i^{\text{th}}$  mass (and no others) in a chain to a unit force, and then measure the resulting displacement of the  $j^{\text{th}}$  mass. Prove that this is the *same* as the displacement of the  $i^{\text{th}}$  mass when the chain is subject to a unit force on the  $j^{\text{th}}$  mass. *Hint:* See Exercise 1.6.20.
- ♣ 6.1.7. Find the displacements  $u_1, u_2, \dots, u_{100}$  of 100 masses connected in a row by identical springs, with spring constant  $c = 1$ . Consider the following three types of force functions: (a) Constant force:  $f_1 = \dots = f_{100} = .01$ ; (b) Linear force:  $f_i = .0002i$ ; (c) Quadratic force:  $f_i = 6 \cdot 10^{-6}i(100 - i)$ . Also consider two different boundary conditions at the bottom: (i) spring 101 connects the last mass to a support; (ii) mass 100 hangs free at the end of the line of springs. Graph the displacements and elongations in all six cases. Discuss your results; in particular, comment on whether they agree with your physical intuition.
- 6.1.8. (a) Suppose you are given three springs with respective stiffnesses  $c = 1$ ,  $c' = 2$ ,  $c'' = 3$ . In what order should you connect them to three masses and a top support so that the bottom mass goes down the farthest under a uniform gravitational force? (b) Answer Exercise 6.1.8 when the springs connect two masses to top and bottom supports.
- ♣ 6.1.9. Generalizing Exercise 6.1.8, suppose you are given  $n$  different springs. (a) In which order should you connect them to  $n$  masses and a top support so that the bottom mass goes down the farthest under a uniform gravitational force? Does your answer depend upon the relative sizes of the spring constants? (b) Answer the same question when the springs connect  $n - 1$  masses to both top and bottom supports.

- 6.1.10. Find the  $LDL^T$  factorization of an  $n \times n$  tridiagonal matrix whose diagonal entries are all equal to 2 and whose sub- and super-diagonal entries are all equal to  $-1$ . *Hint:* Start with the  $3 \times 3$  case (6.13), and then analyze a slightly larger one to spot the pattern.
- ◇ 6.1.11. In a statically indeterminate situation, the equations  $A^T \mathbf{y} = \mathbf{f}$  do not have a unique solution for the internal forces  $\mathbf{y}$  in terms of the external forces  $\mathbf{f}$ . (a) Prove that, nevertheless, if  $C = I$ , the internal forces are the *unique* solution of minimal Euclidean norm, as given by Theorem 4.50. (b) Use this method to directly find the internal force for the system in Example 6.1. Make sure that your values agree with those in the example.

## Positive Definiteness and the Minimization Principle

You may have already observed that the stiffness matrix  $K = A^T C A$  of a mass–spring chain has the form of a Gram matrix, cf. (3.64), for the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w}$  induced by the diagonal matrix of spring stiffnesses. Moreover, since  $A$  has linearly independent columns (which should be checked), and  $C$  is positive definite, Theorem 3.37 tells us that the stiffness matrix is positive definite:  $K > 0$ . In particular, Theorem 3.43 guarantees that  $K$  is nonsingular, and hence the linear system (6.11) has a unique solution  $\mathbf{u} = K^{-1} \mathbf{f}$ . We can therefore conclude that the mass–spring chain assumes a unique equilibrium position under an arbitrary external force. However, one must keep in mind that this is a mathematical result and may not hold in all physical situations. Indeed, we should anticipate that a very large force will take us outside the regime covered by the linear Hooke’s law relation (6.5), and render our simple mathematical model physically irrelevant.

According to Theorem 5.2, when the coefficient matrix of a linear system is positive definite, the equilibrium solution can be characterized by a minimization principle. For mass–spring chains, the quadratic function to be minimized has a physical interpretation: it is the potential energy of the system. Nature is parsimonious with energy, so a physical system seeks out an energy-minimizing equilibrium configuration. Energy minimization principles are of almost universal validity, and can be advantageously used for the construction of mathematical models, as well as their solutions, both analytical and numerical.

The energy function to be minimized can be determined directly from physical principles. For a mass–spring chain, the potential energy of the  $i^{\text{th}}$  mass equals the product of the applied force and the displacement:  $-f_i u_i$ . The minus sign is the result of our convention that a positive displacement  $u_i > 0$  means that the mass has moved down, and hence decreased its potential energy. Thus, the total potential energy due to external forcing on all the masses in the chain is

$$- \sum_{i=1}^n f_i u_i = -\mathbf{u}^T \mathbf{f}.$$

Next, we calculate the internal energy of the system. In a single spring elongated by an amount  $e$ , the work done by the internal forces  $y = ce$  is stored as potential energy, and so is calculated by integrating the force over the elongated distance:

$$\int_0^e y \, de = \int_0^e c e \, de = \frac{1}{2} c e^2.$$

Totaling the contributions from each spring, we find the internal spring energy to be

$$\frac{1}{2} \sum_{i=1}^n c_i e_i^2 = \frac{1}{2} \mathbf{e}^T C \mathbf{e} = \frac{1}{2} \mathbf{u}^T A^T C A \mathbf{u} = \frac{1}{2} \mathbf{u}^T K \mathbf{u},$$

where we used the incidence equation  $\mathbf{e} = A\mathbf{u}$  relating elongation and displacement. Therefore, the total potential energy is

$$p(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T K \mathbf{u} - \mathbf{u}^T \mathbf{f}. \quad (6.16)$$

Since  $K > 0$ , Theorem 5.2 implies that this quadratic function has a unique minimizer that satisfies the equilibrium equation  $K\mathbf{u} = \mathbf{f}$ .

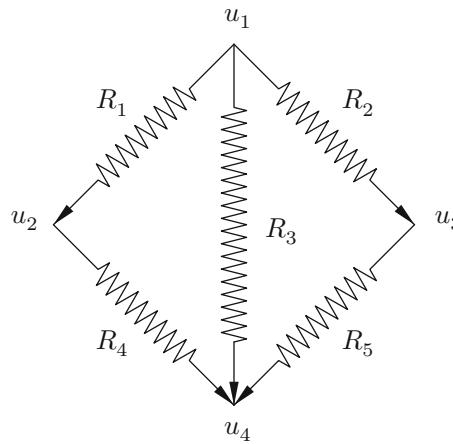
**Example 6.3.** For the three mass chain with two fixed ends described in Example 6.1, the potential energy function (6.16) has the explicit form

$$\begin{aligned} p(\mathbf{u}) &= \frac{1}{2} (u_1 \ u_2 \ u_3) \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} - (u_1 \ u_2 \ u_3) \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \\ &= u_1^2 - u_1 u_2 + u_2^2 - u_2 u_3 + u_3^2 - f_1 u_1 - f_2 u_2 - f_3 u_3, \end{aligned}$$

where  $\mathbf{f} = (f_1, f_2, f_3)^T$  is the external forcing. The minimizer of this particular quadratic function gives the equilibrium displacements  $\mathbf{u} = (u_1, u_2, u_3)^T$  of the three masses.

## Exercises

- 6.1.12. Prove directly that the stiffness matrices in Examples 6.1 and 6.2 are positive definite.
- 6.1.13. Write down the potential energy for the following mass–spring chains with identical unit springs when subject to a uniform gravitational force: (a) three identical masses connected to only a top support. (b) four identical masses connected to top and bottom supports. (c) four identical masses connected only to a top support.
- 6.1.14. (a) Find the total potential energy of the equilibrium configuration of the mass–spring chain in Exercise 6.1.1. (b) Test the minimum principle by substituting three other possible displacements of the masses and checking that they all have larger potential energy.
- 6.1.15. Answer Exercise 6.1.14 for the mass–spring chain in Exercise 6.1.4.
- 6.1.16. Describe the mass–spring chains that gives rise to the following potential energy functions, and find their equilibrium configuration: (a)  $3u_1^2 - 4u_1 u_2 + 3u_2^2 + u_1 - 3u_2$ , (b)  $5u_1^2 - 6u_1 u_2 + 3u_2^2 + 2u_2$ , (c)  $2u_1^2 - 3u_1 u_2 + 4u_2^2 - 5u_2 u_3 + \frac{5}{2}u_3^2 - u_1 - u_2 + u_3$ , (d)  $2u_1^2 - u_1 u_2 + u_2^2 - u_2 u_3 + u_3^2 - u_3 u_4 + 2u_4^2 + u_1 - 2u_3$ .
- 6.1.17. Explain why the columns of the reduced incidence matrices (6.4) and (6.14) are linearly independent.
- 6.1.18. Suppose that when subject to a nonzero external force  $\mathbf{f} \neq \mathbf{0}$ , a mass–spring chain has equilibrium position  $\mathbf{u}^*$ . Prove that the potential energy is strictly negative at equilibrium:  $p(\mathbf{u}^*) < 0$ .
- ◇ 6.1.19. Return to the situation investigated in Exercise 6.1.8. How should you arrange the springs in order to minimize the potential energy in the resulting mass–spring chain?
- 6.1.20. *True or false:* The potential energy function uniquely determines the mass–spring chain.



**Figure 6.5.** A Simple Electrical Network.

## 6.2 Electrical Networks

By an electrical *network*, we mean a collection of (insulated) wires that are joined together at their ends. The junctions connecting the ends of one or more wires are called *nodes*. Mathematically, we can view any such electrical network as a graph, the wires being the edges and the nodes the vertices. As before, to avoid technicalities, we will assume that the underlying graph is *simple*, meaning that there are no loops and at most one edge connecting any two vertices. To begin with, we further assume that there are no electrical devices (batteries, inductors, capacitors, etc.) in the network, and so the only impediments to the current flowing through the network are the resistances in the wires. As we shall see, resistance (or rather its reciprocal) plays a very similar role to that of spring stiffness. Thus, the network corresponds to a *weighted graph* in which the weight of an edge is the number representing the resistance of the corresponding wire. We shall feed a current into the network at one or more of the nodes, and would like to determine how the induced current flows through the wires. The basic equations governing the equilibrium voltages and currents in such a network follow from the three fundamental laws of electricity, named after the pioneering nineteenth-century German physicists Gustav Kirchhoff and Georg Ohm, two of the founders of electric circuit theory, [58].

*Voltage* is defined as the electromotive force that moves electrons through a wire. An individual wire's voltage is determined by the difference in the voltage potentials at its two ends — just as the gravitational force on a mass is induced by a difference in gravitational potential. To quantify voltage, we need to fix an orientation for the wire. A positive voltage will mean that the electrons move in the chosen direction, while a negative voltage causes them to move in reverse. The original choice of orientation is arbitrary, but once assigned will pin down the sign conventions to be used by voltages, currents, etc. To this end, we draw a digraph to represent the network, whose edges represent wires and whose vertices represent nodes. Each edge is assigned an orientation that indicates the wire's starting and ending nodes. A simple example consisting of five wires joined at four different nodes can be seen in Figure 6.5. The arrows indicate the selected directions for the wires, the wavy lines are the standard electrical symbols for resistance, while the resistances provide the edge weights in the resulting *weighted digraph*.

In an electrical network, each node will have a voltage potential, denoted by  $u_i$ . If wire  $k$  starts at node  $i$  and ends at node  $j$  under its assigned orientation, then its voltage  $v_k$

equals the difference between the voltage potentials at its ends:

$$v_k = u_i - u_j. \quad (6.17)$$

Note that  $v_k > 0$  if  $u_i > u_j$ , indicating that the electrons flow from the starting node  $i$  to the ending node  $j$ . In our particular illustrative example, the five wires have respective voltages

$$v_1 = u_1 - u_2, \quad v_2 = u_1 - u_3, \quad v_3 = u_1 - u_4, \quad v_4 = u_2 - u_4, \quad v_5 = u_3 - u_4.$$

Let us rewrite this linear system of equations in vector form

$$\mathbf{v} = A\mathbf{u}, \quad (6.18)$$

where

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \quad (6.19)$$

The alert reader will recognize the *incidence matrix* (2.46) for the digraph defined by the network. This is true in general — *the voltages along the wires of an electrical network are related to the potentials at the nodes by a linear system of the form (6.18), in which  $A$  is the incidence matrix of the network digraph*. The rows of the incidence matrix are indexed by the wires, and the columns by the nodes. Each row of the matrix  $A$  has a single  $+1$  in the column indexed by the starting node of the associated wire, and a single  $-1$  in the column of the ending node.

*Kirchhoff's Voltage Law* states that the sum of the voltages around each closed circuit in the network is zero. For example, in the network under consideration, summing the voltages around the left-hand triangular circuit gives

$$v_1 + v_4 - v_3 = (u_1 - u_2) + (u_2 - u_4) - (u_1 - u_4) = 0.$$

Note that  $v_3$  appears with a minus sign, since we must traverse wire 3 in the opposite direction to its assigned orientation when going around the circuit in the counterclockwise direction. The voltage law is a direct consequence of (6.18). Indeed, as discussed in Section 2.6, the circuits can be identified with vectors  $\ell \in \text{coker } A = \ker A^T$  in the cokernel of the incidence matrix, and so

$$\ell \cdot \mathbf{v} = \ell^T \mathbf{v} = \ell^T A \mathbf{u} = 0. \quad (6.20)$$

Therefore, orthogonality of the voltage vector  $\mathbf{v}$  to the circuit vector  $\ell$  is the mathematical formalization of Kirchhoff's Voltage Law.

Given a prescribed set of voltages  $\mathbf{v}$  along the wires, can one find corresponding voltage potentials  $\mathbf{u}$  at the nodes? To answer this question, we need to solve  $\mathbf{v} = A\mathbf{u}$ , which requires  $\mathbf{v} \in \text{img } A$ . According to the Fredholm Alternative Theorem 4.46, the necessary and sufficient condition for this to hold is that  $\mathbf{v}$  be orthogonal to  $\text{coker } A$ . Theorem 2.53 says that the cokernel of an incidence matrix is spanned by the circuit vectors, and so  $\mathbf{v}$  is a possible set of voltages if and only if  $\mathbf{v}$  is orthogonal to all the circuit vectors  $\ell \in \text{coker } A$ , i.e., the Voltage Law is necessary and sufficient for the given voltages to be physically realizable in the network.

Kirchhoff's Law is related to the topology of the network — how the different wires are connected together. *Ohm's Law* is a constitutive relation, indicating what the wires are

made of. The resistance along a wire (including any added resistors) prescribes the relation between voltage and current or the rate of flow of electric charge. The law reads

$$v_k = R_k y_k, \quad (6.21)$$

where  $v_k$  is the voltage,  $R_k$  is the resistance, and  $y_k$  (often denoted by  $I_k$  in the engineering literature) denotes the current along wire  $k$ . Thus, for a fixed voltage, the larger the resistance of the wire, the smaller the current that flows through it. The direction of the current is also prescribed by our choice of orientation of the wire, so that  $y_k > 0$  if the current is flowing from the starting to the ending node. We combine the individual equations (6.21) into a single vector equation

$$\mathbf{v} = R \mathbf{y}, \quad (6.22)$$

where the *resistance matrix*  $R = \text{diag}(R_1, \dots, R_n) > 0$  is diagonal and positive definite. We shall, in analogy with (6.6), replace (6.22) by the inverse relationship

$$\mathbf{y} = C \mathbf{v}, \quad (6.23)$$

where  $C = R^{-1}$  is the *conductance matrix*, again diagonal, positive definite, whose entries are the *conductances*  $c_k = 1/R_k$  of the wires. For the particular network in [Figure 6.5](#),

$$C = \begin{pmatrix} c_1 & 0 & 0 & 0 & 0 \\ 0 & c_2 & 0 & 0 & 0 \\ 0 & 0 & c_3 & 0 & 0 \\ 0 & 0 & 0 & c_4 & 0 \\ 0 & 0 & 0 & 0 & c_5 \end{pmatrix} = \begin{pmatrix} 1/R_1 & 0 & 0 & 0 & 0 \\ 0 & 1/R_2 & 0 & 0 & 0 \\ 0 & 0 & 1/R_3 & 0 & 0 \\ 0 & 0 & 0 & 1/R_4 & 0 \\ 0 & 0 & 0 & 0 & 1/R_5 \end{pmatrix} = R^{-1}. \quad (6.24)$$

Finally, we stipulate that electric current is not allowed to accumulate at any node, i.e., every electron that arrives at a node must leave along one of the wires. Let  $y_k, y_l, \dots, y_m$  denote the currents along all the wires  $k, l, \dots, m$  that meet at node  $i$  in the network, and  $f_i$  an external current source, if any, applied at node  $i$ . *Kirchhoff's Current Law* requires that the net current leaving the node along the wires equals the external current coming into the node, and so

$$\pm y_k \pm y_l \pm \dots \pm y_m = f_i. \quad (6.25)$$

Each  $\pm$  sign is determined by the orientation of the wire, with  $+$  if node  $i$  is its starting node and  $-$  if it is its ending node.

In our particular example, suppose that we send a 1 amp current source into the first node. Then Kirchhoff's Current Law requires

$$y_1 + y_2 + y_3 = 1, \quad -y_1 + y_4 = 0, \quad -y_2 + y_5 = 0, \quad -y_3 - y_4 - y_5 = 0,$$

the four equations corresponding to the four nodes in our network. The vector form of this linear system is

$$A^T \mathbf{y} = \mathbf{f}, \quad (6.26)$$

where  $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5)^T$  are the currents along the five wires, and  $\mathbf{f} = (1, 0, 0, 0, 0)^T$  represents the current sources at the four nodes. The coefficient matrix

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & -1 \end{pmatrix} \quad (6.27)$$

is the *transpose* of the incidence matrix (6.19). As in the mass–spring chain, this is a remarkable general fact, which follows directly from Kirchhoff’s two laws. *The coefficient matrix for the Current Law is the transpose of the incidence matrix for the Voltage Law.*

Let us assemble the full system of equilibrium equations (6.18, 23, 26):

$$\mathbf{v} = A\mathbf{u}, \quad \mathbf{y} = C\mathbf{v}, \quad \mathbf{f} = A^T\mathbf{y}. \quad (6.28)$$

Remarkably, we arrive at a system of linear relations that has an identical form to the mass–spring chain system (6.10), albeit with different physical quantities and different coefficient matrices. As before, they combine into a single linear system

$$K\mathbf{u} = \mathbf{f}, \quad \text{where} \quad K = A^T C A \quad (6.29)$$

is known as the *resistivity matrix* associated with the network. In our particular example, combining (6.19, 24, 27) produces the resistivity matrix

$$K = A^T C A = \begin{pmatrix} c_1 + c_2 + c_3 & -c_1 & -c_2 & -c_3 \\ -c_1 & c_1 + c_4 & 0 & -c_4 \\ -c_2 & 0 & c_2 + c_5 & -c_5 \\ -c_3 & -c_4 & -c_5 & c_3 + c_4 + c_5 \end{pmatrix}, \quad (6.30)$$

whose entries depend on the conductances of the five wires in the network.

**Remark.** There is a simple pattern to the resistivity matrix, evident in (6.30). The diagonal entries  $k_{ii}$  equal the sum of the conductances of all the wires having node  $i$  at one end. The non-zero off-diagonal entries  $k_{ij}$ ,  $i \neq j$ , equal  $-c_k$ , the conductance of the wire<sup>†</sup> joining node  $i$  to node  $j$ , while  $k_{ij} = 0$  if there is no wire joining the two nodes.

Consider the case in which all the wires in our network have equal unit resistance, and so  $c_k = 1/R_k = 1$  for  $k = 1, \dots, 5$ . Then the resistivity matrix is

$$K = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}. \quad (6.31)$$

However, when trying to solve the linear system (6.29), we run into an immediate difficulty: *there is no solution!* The matrix (6.31) is *not* positive definite — it is a singular matrix. Moreover, the particular current source vector  $\mathbf{f} = (1, 0, 0, 0)^T$  does not lie in the image of  $K$ . Something is clearly amiss.

Before getting discouraged, let us sit back and use a little physical intuition. We are trying to put a 1 amp current into the network at node 1. Where can the electrons go? The answer is nowhere — they are all trapped in the network and, as they accumulate, something drastic will happen — sparks will fly! This is clearly an unstable situation, and so the fact that the equilibrium equations do not have a solution is trying to tell us that the physical system cannot remain in a steady state. The physics rescues the mathematics, or, vice versa, the mathematics elucidates the underlying physical processes.

In order to achieve equilibrium in an electrical network, we must remove as much current as we put in. Thus, if we feed a 1 amp current into node 1, then we must extract a total of

---

<sup>†</sup> This assumes that there is only one wire joining the two nodes.

1 amp's worth of current from the other nodes. In other words, the sum of all the external current sources must vanish:

$$f_1 + f_2 + \cdots + f_n = 0,$$

and so there is no net current being fed into the network. Suppose we also extract a 1 amp current from node 4; then the modified current source vector  $\mathbf{f} = (1, 0, 0, -1)^T$  indeed lies in the image of  $K$ , as you can check, and the equilibrium system (6.29) has a solution.

This is all well and good, but we are not out of the woods yet. As we know, if a linear system has a singular coefficient matrix, then either it has no solutions — the case we already rejected — or it has infinitely many solutions — the case we are considering now. In the particular network under consideration, the general solution to the linear system

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

is found by Gaussian Elimination:

$$\mathbf{u} = \begin{pmatrix} \frac{1}{2} + t \\ \frac{1}{4} + t \\ \frac{1}{4} + t \\ t \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad (6.32)$$

where  $t = u_4$  is the free variable. The resulting nodal voltage potentials

$$u_1 = \frac{1}{2} + t, \quad u_2 = \frac{1}{4} + t, \quad u_3 = \frac{1}{4} + t, \quad u_4 = t,$$

depend on a free parameter  $t$ .

The ambiguity arises because voltage potential is a mathematical abstraction that cannot be measured directly; only relative potential differences have physical import. To resolve the inherent ambiguity, we need to assign a baseline value for the voltage potentials. In terrestrial electricity, the Earth is assumed to have zero potential. Specifying a particular node to have zero potential is physically equivalent to grounding that node. For our example, suppose we ground node 4 by setting  $u_4 = 0$ . This fixes the free variable  $t = 0$  in our solution (6.32), and so uniquely specifies all the other voltage potentials:  $u_1 = \frac{1}{2}$ ,  $u_2 = \frac{1}{4}$ ,  $u_3 = \frac{1}{4}$ ,  $u_4 = 0$ .

On the other hand, even without specification of a baseline potential level, the corresponding physical voltages and currents along the wires are uniquely specified. In our example, computing  $\mathbf{y} = \mathbf{v} = A\mathbf{u}$  gives

$$y_1 = v_1 = \frac{1}{4}, \quad y_2 = v_2 = \frac{1}{4}, \quad y_3 = v_3 = \frac{1}{2}, \quad y_4 = v_4 = \frac{1}{4}, \quad y_5 = v_5 = \frac{1}{4},$$

independent of the value of  $t$  in (6.32). Thus, the nonuniqueness of the voltage potential solution  $\mathbf{u}$  is an inessential feature. All physical quantities that we can measure — currents and voltages — are uniquely specified by the solution to the equilibrium system.

**Remark.** Although they have no real physical meaning, we *cannot* dispense with the nonmeasurable (and nonunique) voltage potentials  $\mathbf{u}$ . Most networks are *statically indeterminate*, since their incidence matrices are rectangular and hence not invertible, so the linear system  $A^T\mathbf{y} = \mathbf{f}$  cannot be solved directly for the currents in terms of the voltage

sources since the system does not have a unique solution. Only by first solving the full equilibrium system (6.29) for the potentials, and then using the relation  $\mathbf{y} = CA\mathbf{u}$  between the potentials and the currents, can we determine their actual values.

Let us analyze what is going on in the context of our general mathematical framework. Proposition 3.36 says that the resistivity matrix  $K = A^T CA$  is a positive semi-definite Gram matrix, which is positive definite (and hence nonsingular) if and only if  $A$  has linearly independent columns, or, equivalently,  $\ker A = \{\mathbf{0}\}$ . But Proposition 2.51 says that the incidence matrix  $A$  of a directed graph *never* has a trivial kernel. Therefore, the resistivity matrix  $K$  is only positive semi-definite, and hence singular. If the network is connected, then  $\ker A = \ker K = \text{coker } K$  is one-dimensional, spanned by the vector  $\mathbf{z} = (1, 1, 1, \dots, 1)^T$ . According to the Fredholm Alternative Theorem 4.46, the fundamental network equation  $K\mathbf{u} = \mathbf{f}$  has a solution if and only if  $\mathbf{f}$  is orthogonal to  $\text{coker } K$ , and so the current source vector must satisfy

$$\mathbf{z} \cdot \mathbf{f} = f_1 + f_2 + \dots + f_n = 0, \quad (6.33)$$

as we already observed. Therefore, the linear algebra reconfirms our physical intuition: a connected network admits an equilibrium configuration, obtained by solving (6.29), if and only if the nodal current sources add up to zero, i.e., there is no net influx of current into the network.

Grounding one of the nodes is equivalent to nullifying the value of its voltage potential:  $u_i = 0$ . This variable is now fixed, and can be safely eliminated from our system. To accomplish this, we let  $A^*$  denote the  $m \times (n - 1)$  matrix obtained by deleting the  $i^{\text{th}}$  column from  $A$ . For example, grounding node 4 in our sample network, so  $u_4 = 0$ , allows us to erase the fourth column of the incidence matrix (6.19), leading to the *reduced incidence matrix*

$$A^* = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (6.34)$$

The key observation is that  $A^*$  has trivial kernel,  $\ker A^* = \{\mathbf{0}\}$ , and therefore the reduced network resistivity matrix

$$K^* = (A^*)^T C A^* = \begin{pmatrix} c_1 + c_2 + c_3 & -c_1 & -c_2 \\ -c_1 & c_1 + c_4 & 0 \\ -c_2 & 0 & c_2 + c_5 \end{pmatrix} \quad (6.35)$$

is positive definite. Note that we can obtain  $K^*$  directly from  $K$  in (6.30) by deleting both its fourth row and fourth column. Let  $\mathbf{f}^* = (1, 0, 0)^T$  denote the reduced current source vector obtained by deleting the fourth entry from  $\mathbf{f}$ . Then the reduced linear system is

$$K^* \mathbf{u}^* = \mathbf{f}^*, \quad (6.36)$$

where  $\mathbf{u}^* = (u_1, u_2, u_3)^T$  is the reduced voltage potential vector. Positive definiteness of  $K^*$  implies that (6.36) has a unique solution  $\mathbf{u}^*$ , from which we can reconstruct the voltages  $\mathbf{v} = A^* \mathbf{u}^*$  and currents  $\mathbf{y} = C \mathbf{v} = C A^* \mathbf{u}^*$  along the wires. In our example, if all the wires have unit resistance, then the reduced system (6.36) is

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

and has unique solution  $\mathbf{u}^* = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})^T$ . The voltage potentials are

$$u_1 = \frac{1}{2}, \quad u_2 = \frac{1}{4}, \quad u_3 = \frac{1}{4}, \quad u_4 = 0,$$

and correspond to the earlier solution (6.32) when  $t = 0$ . The corresponding voltages and currents along the wires are the same as before.

**Remark.** When  $C = I$ , the matrix  $K = A^T A$  constructed from the incidence matrix of a directed graph is known in the mathematical literature as the *graph Laplacian* associated with the graph. The graph Laplacian matrix can be easily constructed directly: its rows and columns are indexed by the vertices. The diagonal entry  $k_{ii}$  equals the degree of the  $i^{\text{th}}$  vertex, meaning the number of edges that have vertex  $i$  as one of their endpoints. The off-diagonal entries  $k_{ij}$  are equal to  $-1$  if there is an edge connecting vertices  $i$  and  $j$  and  $0$  otherwise. This is often written as  $K = D - J$ , where  $D$  is the diagonal *degree matrix*, whose diagonal entries are the degrees of the nodes, and  $J$  is the symmetric *adjacency matrix*, which contains a  $1$  in every off-diagonal entry corresponding to two adjacent nodes, that is two nodes connected by a single edge; all other entries are  $0$ . Observe that the graph Laplacian is independent of the direction assigned to the edges; it depends only on the underlying graph. The term “Laplacian” is used because this matrix represents the discrete analogue of the *Laplacian* differential operator, described in Examples 7.36 and 7.52 below. In particular, if the graph comes from an  $n$ -dimensional square grid, the corresponding graph Laplacian coincides with the standard finite difference numerical discretization of the Laplacian differential operator.

## Batteries, Power, and the Electrical–Mechanical Correspondence

So far, we have considered only the effect of current sources at the nodes. Suppose now that the network contains one or more batteries. Each *battery* serves as a voltage source along a wire, and we let  $b_k$  denote the voltage of a battery connected to wire  $k$ . The sign of  $b_k$  indicates the relative orientation of the battery’s terminals with respect to the wire, with  $b_k > 0$  if the current produced by the battery runs in the same direction as our chosen orientation of the wire. The battery’s voltage is included in the voltage balance equation (6.17):

$$v_k = u_i - u_j + b_k.$$

The corresponding vector equation (6.18) becomes

$$\mathbf{v} = A\mathbf{u} + \mathbf{b}, \tag{6.37}$$

where  $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$  is the *battery vector*, whose entries are indexed by the wires. (If there is no battery on wire  $k$ , the corresponding entry is  $b_k = 0$ .) The remaining two equations are as before, so  $\mathbf{y} = C\mathbf{v}$  are the currents in the wires, and, in the absence of external current sources, Kirchhoff’s Current Law implies  $A^T \mathbf{y} = \mathbf{0}$ . Using the modified formula (6.37) for the voltages, these combine into the following equilibrium system:

$$K\mathbf{u} = A^T C A \mathbf{u} = -A^T C \mathbf{b}. \tag{6.38}$$

**Remark.** Interestingly, the voltage potentials satisfy the weighted normal equations (5.36) that characterize the least squares solution to the system  $A\mathbf{u} = -\mathbf{b}$  for the weighted norm

$$\|\mathbf{v}\|^2 = \mathbf{v}^T C \mathbf{v} \tag{6.39}$$

determined by the network’s conductance matrix  $C$ . It is a striking fact that Nature solves a least squares problem in order to make the weighted norm of the voltages  $\mathbf{v}$  as small as possible. A similar remark holds for the mass–spring chains considered above.

Batteries have exactly the same effect on the voltage potentials as if we imposed the current source vector

$$\mathbf{f} = -A^T C \mathbf{b}. \quad (6.40)$$

Namely, placing a battery of voltage  $b_k$  on wire  $k$  is exactly the same as introducing additional current sources of  $-c_k b_k$  at the starting node and  $c_k b_k$  at the ending node. Note that the induced current vector  $\mathbf{f} \in \text{coimg } A = \text{img } K$  (see Exercise 3.4.32) continues to satisfy the network constraint (6.33). Conversely, a system of allowed current sources  $\mathbf{f} \in \text{img } K$  has the same effect as any collection of batteries  $\mathbf{b}$  that satisfies (6.40).

In the absence of external current sources, a network with batteries always admits a solution for the voltage potentials and currents. Although the currents are uniquely determined, the voltage potentials are not. As before, to eliminate the ambiguity, we can ground one of the nodes and use the reduced incidence matrix  $A^*$  and reduced current source vector  $\mathbf{f}^*$  obtained by eliminating the column, respectively entry, corresponding to the grounded node. The details are left to the interested reader.

**Example 6.4.** Consider an electrical network running along the sides of a cube, where each wire contains a 2 ohm resistor and there is a 9 volt battery source on one wire. The problem is to determine how much current flows through the wire directly opposite the battery. Orienting the wires and numbering them as indicated in [Figure 6.6](#), the incidence matrix is

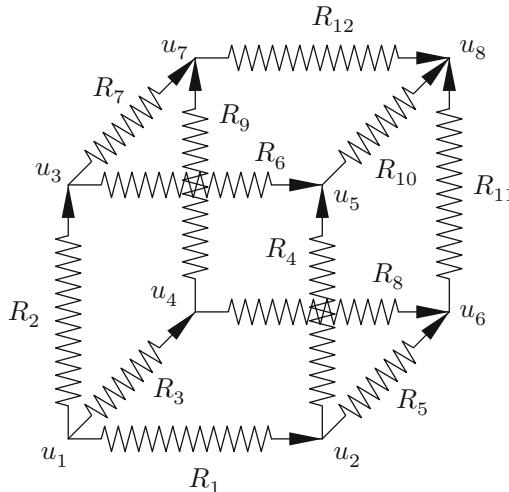
$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 &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

We connect the battery along wire 1 and measure the resulting current along wire 12. To avoid the ambiguity in the voltage potentials, we ground the last node and erase the final column from  $A$  to obtain the reduced incidence matrix  $A^*$ . Since the resistance matrix  $R$  has all 2's along the diagonal, the conductance matrix is  $C = \frac{1}{2} I$ . Therefore, the network resistivity matrix is one-half the cubical graph Laplacian:

$$K^* = (A^*)^T C A^* = \frac{1}{2} (A^*)^T A^* = \frac{1}{2} \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & -1 & -1 & 0 \\ -1 & 0 & 3 & 0 & -1 & 0 & -1 \\ -1 & 0 & 0 & 3 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & 3 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 3 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 3 \end{pmatrix}.$$

Alternatively, it can be found by eliminating the final row and column, representing the grounded node, from the graph Laplacian matrix constructed by the above recipe. The reduced current source vector

$$\mathbf{b} = (9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^T$$



**Figure 6.6.** Cubical Electrical Network.

corresponding to the battery situated on the first wire is

$$\mathbf{f}^* = -(A^*)^T C \mathbf{b} = \left( -\frac{9}{2}, \frac{9}{2}, 0, 0, 0, 0, 0, 0 \right)^T.$$

Solving the resulting linear system  $K^* \mathbf{u}^* = \mathbf{f}^*$  by Gaussian Elimination yields the voltage potentials

$$\mathbf{u}^* = \left( -3, \frac{9}{4}, -\frac{9}{8}, -\frac{9}{8}, \frac{3}{8}, \frac{3}{8}, -\frac{3}{4} \right)^T.$$

Thus, the induced currents along the sides of the cube are

$$\mathbf{y} = C \mathbf{v} = \frac{1}{2} (A^* \mathbf{u}^* + \mathbf{b}) = \left( \frac{15}{8}, -\frac{15}{16}, -\frac{15}{16}, \frac{15}{16}, \frac{15}{16}, -\frac{3}{4}, -\frac{3}{16}, -\frac{3}{4}, -\frac{3}{16}, \frac{3}{16}, \frac{3}{16}, -\frac{3}{8} \right)^T.$$

In particular, the current on the wire that is opposite the battery is  $y_{12} = -\frac{3}{8}$ , flowing in the opposite direction to its orientation. The largest current flows through the battery wire, while wires 7, 9, 10, 11 transmit the least.

As with a mass–spring chain, the voltage potentials in such a resistive electrical network can be characterized by a minimization principle. The *power* in a single conducting wire is defined as the product of its current  $y_j$  and voltage  $v_j$ ,

$$P_j = y_j v_j = R_j y_j^2 = c_j v_j^2, \quad (6.41)$$

where  $R_j$  is the resistance,  $c_j = 1/R_j$  the conductance, and we are using Ohm's Law (6.21) to relate voltage and current. Physically, the power quantifies the rate at which electrical energy is converted into heat by the wire's resistance. Summing over all wires in the system, the internal power<sup>†</sup> of the network

$$P_{int} = \sum_j P_j = \sum_j c_j v_j^2 = \|\mathbf{v}\|^2$$

is identified as the square of the weighted norm (6.39).

---

<sup>†</sup> So far, we have not considered the effect of batteries or current sources on the network.

### The Electrical–Mechanical Correspondence

Structures	Variables	Networks
Displacements	$\mathbf{u}$	Voltage potentials
Prestressed bars/springs	$\mathbf{b}$	Batteries
Elongations <sup>†</sup>	$\mathbf{v} = A\mathbf{u} + \mathbf{b}$	Voltages
Spring stiffnesses	$C$	Conductivities
Internal Forces	$\mathbf{y} = C\mathbf{v}$	Currents
External forcing	$\mathbf{f} = A^T\mathbf{y}$	Current sources
Stiffness matrix	$K = A^TCA$	Resistivity matrix
Potential energy	$p(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T K \mathbf{u} - \mathbf{u}^T \mathbf{f}$	$\frac{1}{2} \times \text{Power}$

Consider a network that contains batteries, but no external current sources. Summing over all the wires in the network, the total power due to internal and external sources can be identified as the product of the current and voltage vectors:

$$\begin{aligned} P = y_1 v_1 + \cdots + y_m v_m &= \mathbf{y}^T \mathbf{v} = \mathbf{v}^T C \mathbf{v} = (A\mathbf{u} + \mathbf{b})^T C (A\mathbf{u} + \mathbf{b}) \\ &= \mathbf{u}^T A^T C A \mathbf{u} + 2\mathbf{u}^T A^T C \mathbf{b} + \mathbf{b}^T C \mathbf{b}, \end{aligned}$$

and is thus a quadratic function of the voltage potentials, which we rewrite in our usual form<sup>‡</sup>

$$\frac{1}{2}P = p(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T K \mathbf{u} - \mathbf{u}^T \mathbf{f} + c, \quad (6.42)$$

where  $K = A^T C A$  is the network resistivity matrix, while  $\mathbf{f} = -A^T C \mathbf{b}$  are the equivalent current sources at the nodes (6.40) that correspond to the batteries. The last term  $c = \frac{1}{2}\mathbf{b}^T C \mathbf{b}$  is one-half the internal power of the batteries, and is not affected by the currents/voltages in the wires. In deriving (6.42), we have ignored external current sources at the nodes. By the preceding discussion, external current sources can be viewed as an equivalent collection of batteries, and so contribute to the linear terms  $\mathbf{u}^T \mathbf{f}$  in the power, which will then represent the combined effect of all batteries and external current sources.

In general, the resistivity matrix  $K$  is only positive semi-definite, and so the quadratic power function (6.42) does not, in general, possess a minimizer. As argued above, to ensure equilibrium, we need to ground one or more of the nodes. The resulting reduced power function

$$p^*(\mathbf{u}^*) = \frac{1}{2}(\mathbf{u}^*)^T K^* \mathbf{u}^* - (\mathbf{u}^*)^T \mathbf{f}^*, \quad (6.43)$$

has a positive definite coefficient matrix:  $K^* > 0$ . Its unique minimizer is the voltage potential  $\mathbf{u}^*$  that solves the reduced linear system (6.36). We conclude that the electrical network adjusts itself so as to *minimize the power or total energy loss* throughout the network. As in mechanics, Nature solves a minimization problem in an effort to conserve energy.

<sup>†</sup> Here, we use  $\mathbf{v}$  instead of  $\mathbf{e}$  to represent elongation.

<sup>‡</sup> For alternating currents, the power is reduced by a factor of  $\frac{1}{2}$ , so  $p(\mathbf{u})$  equals the power.

We have now discovered the remarkable correspondence between the equilibrium equations for electrical networks (6.10) and those of mass–spring chains (6.28). This *Electrical–Mechanical Correspondence* is summarized in the above table. In the following section, we will see that the analogy extends to more general mechanical structures.

## Exercises

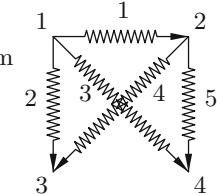
6.2.1. Draw the electrical networks corresponding to the following incidence matrices.

$$(a) \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 1 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 & 0 \end{pmatrix},$$

$$(d) \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix}, \quad (e) \begin{pmatrix} 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

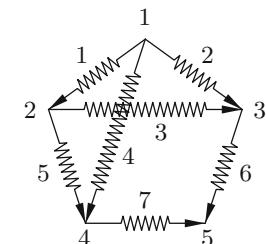
6.2.2. Suppose that all wires in the illustrated network have unit resistivity.

- (a) Write down the incidence matrix  $A$ . (b) Write down the equilibrium system for the network when node 4 is grounded and there is a current source of magnitude 3 at node 1. (c) Solve the system for the voltage potentials at the ungrounded nodes. (d) If you connect a light bulb to the network, which wire should you connect it to so that it shines the brightest?

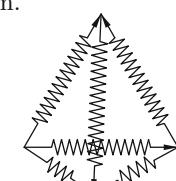


6.2.3. What happens in the network in Figure 6.5 if we ground both nodes 3 and 4? Set up and solve the system and compare the currents for the two cases.

6.2.4. (a) Write down the incidence matrix  $A$  for the illustrated electrical network. (b) Suppose all the wires contain unit resistors, except for  $R_4 = 2$ . Let there be a unit current source at node 1, and assume node 5 is grounded. Find the voltage potentials at the nodes and the currents through the wires. (c) Which wire would shock you the most?



6.2.5. Answer Exercise 6.2.4 if, instead of the current source, you put a 1.5 volt battery on wire 1.



♣ 6.2.6. Consider an electrical network running along the sides of a tetrahedron. Suppose that each wire contains a 3 ohm resistor and there is a 10 volt battery source on one wire. Determine how much current flows through the wire directly opposite the battery.

♣ 6.2.7. Now suppose that each wire in the tetrahedral network in Exercise 6.2.6 contains a 1 ohm resistor and there are two 5 volt battery sources located on two non-adjacent wires. Determine how much current flows through the wires in the network.

♣ 6.2.8. (a) How do the currents change if the resistances in the wires in the cubical network in Example 6.4 are all equal to 1 ohm? (b) What if wire  $k$  has resistance  $R_k = k$  ohms?

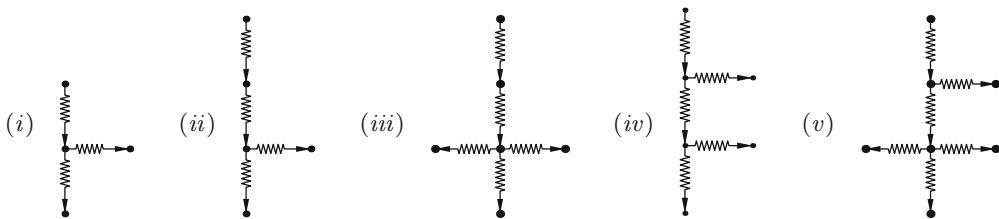
♣ 6.2.9. Suppose you are given six resistors with respective resistances 1, 2, 3, 4, 5, and 6. How should you connect them in a tetrahedral network (one resistor per wire) so that a light bulb on the wire opposite the battery burns the brightest?

- ♣ 6.2.10. The nodes in an electrical network lie on the vertices  $(\frac{i}{n}, \frac{j}{n})$  for  $-n \leq i, j \leq n$  in a square grid centered at the origin; the wires run along the grid lines. The boundary nodes, when  $x$  or  $y = \pm 1$ , are all grounded. A unit current source is introduced at the origin.  
 (a) Compute the potentials at the nodes and currents along the wires for  $n = 2, 3, 4$ .  
 (b) Investigate and compare the solutions for large  $n$ , i.e., as the grid size becomes small.  
 Do you detect any form of limiting behavior?

6.2.11. Show that, in a network with all unit resistors, the currents  $\mathbf{y}$  can be characterized as the unique solution to the Kirchhoff equations  $A^T \mathbf{y} = \mathbf{f}$  of minimum Euclidean norm.

6.2.12. *True or false:* (a) The nodal voltage potentials in a network with batteries  $\mathbf{b}$  are the same as in the same network with the current sources  $\mathbf{f} = -A^T C \mathbf{b}$ . (b) Are the currents the same?

6.2.13. (a) Assuming all wires have unit resistance, find the voltage potentials at all the nodes and the currents along the wires of the following trees when the bottom node is grounded and a unit current source is introduced at the top node.



(b) Can you make any general predictions about electrical currents in trees?

6.2.14. A node in a tree is called *terminating* if it has only one edge. Repeat the preceding exercise when all terminating nodes except for the top one are grounded.

6.2.15. Suppose the graph of an electrical network is a tree, as in Exercise 2.6.9. Show that if one of the nodes in the tree is grounded, the system is statically determinate.

6.2.16. Suppose two wires in a network join the *same* pair of nodes. Explain why their effect on the rest of the network is the same as a single wire whose conductance  $c = c_1 + c_2$  is the sum of the individual conductances. How are the resistances related?

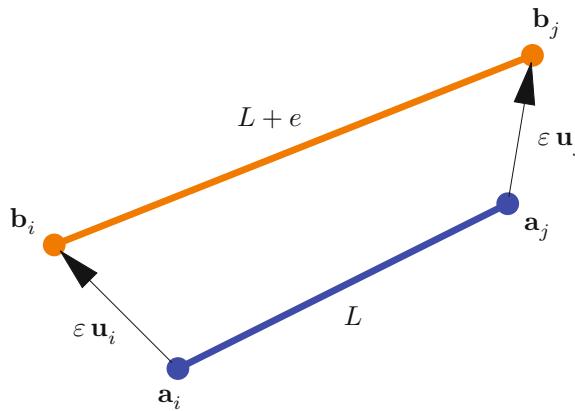
6.2.17. (a) Write down the equilibrium equations for a network that contains both batteries and current sources. (b) Formulate a general superposition principle for such situations.  
 (c) Write down a formula for the power in the network.

◊ 6.2.18. Prove that the voltage potential at node  $i$  due to a unit current source at node  $j$  is the *same* as the voltage potential at node  $j$  due to a unit current source at node  $i$ . Can you give a physical explanation of this *reciprocity relation*?

6.2.19. What is the analogue of condition (6.33) for a disconnected graph?

### 6.3 Structures

A *structure* (sometimes called a *truss*) is a mathematical idealization of a framework for a building. Think of a radio tower or a skyscraper when just the I-beams are connected — before the walls, floors, ceilings, roof, and ornamentation are added. An ideal structure is constructed of elastic bars connected at *joints*. By a *bar*, we mean a straight, rigid rod that can be (slightly) elongated, but not bent. (Beams, which are allowed to bend, are more complicated and are modeled by boundary value problems for ordinary and partial differential equations, [61, 79]. See also our discussion of splines in Section 5.5.) When a bar is stretched, it obeys Hooke's law — at least in the linear regime we are modeling



**Figure 6.7.** Displacement of a Bar.

— and so, for all practical purposes, behaves like a spring with a very large stiffness. As a result, a structure can be regarded as a two- or three-dimensional generalization of a mass–spring chain.

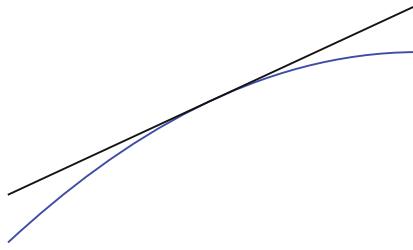
The joints will allow the bar to rotate in any direction. Of course, this is an idealization; in a building, the rivets and bolts will (presumably) prevent rotation to a significant degree. However, under moderate stress — for example, if the wind is blowing on our skyscraper, the bolts can be expected only to keep the structure connected, and the resulting motions will induce stresses on the joints that must be taken into account when designing the structure. Of course, under extreme stress, the structure will fall apart — a disaster that its designers must avoid. The purpose of this section is to derive conditions that will guarantee that a structure is rigidly stable under moderate forcing, or, alternatively, help us to understand the processes that might lead to its collapse.

The first order of business is to understand how an individual bar reacts to motion. We have already encountered the basic idea in our treatment of springs. The key complication here is that the ends of the bar are not restricted to a single direction of motion, but can move in either two- or three-dimensional space. We use  $d$  to denote the dimension of the underlying space. In the  $d = 1$ -dimensional case, the structure reduces to a mass–spring chain that we analyzed in Section 6.1. Here we concentrate on structures in  $d = 2$  and 3 dimensions.

Consider an unstressed bar with one end at position  $\mathbf{a}_1 \in \mathbb{R}^d$  and the other end at position  $\mathbf{a}_2 \in \mathbb{R}^d$ . In  $d = 2$  dimensions, we write  $\mathbf{a}_i = (a_i, b_i)^T$ , while in  $d = 3$ -dimensional space,  $\mathbf{a}_i = (a_i, b_i, c_i)^T$ . The length of the bar is  $L = \|\mathbf{a}_1 - \mathbf{a}_2\|$ , where we use the standard Euclidean norm to measure distance on  $\mathbb{R}^d$  throughout this section.

Suppose we move the ends of the bar a little, sending  $\mathbf{a}_i$  to  $\mathbf{b}_i = \mathbf{a}_i + \varepsilon \mathbf{u}_i$  and, simultaneously,  $\mathbf{a}_j$  to  $\mathbf{b}_j = \mathbf{a}_j + \varepsilon \mathbf{u}_j$ , moving the blue bar in Figure 6.7 to the displaced orange bar. The vectors  $\mathbf{u}_i, \mathbf{u}_j \in \mathbb{R}^d$  indicate the respective directions of displacement of the two ends, and we use  $\varepsilon$  to represent the relative magnitude of the displacement. How much has this motion stretched the bar? Since we are assuming that the bar can't bend, the length of the displaced bar is

$$\begin{aligned} L + e &= \|\mathbf{b}_i - \mathbf{b}_j\| = \|(\mathbf{a}_i + \varepsilon \mathbf{u}_i) - (\mathbf{a}_j + \varepsilon \mathbf{u}_j)\| = \|(\mathbf{a}_i - \mathbf{a}_j) + \varepsilon (\mathbf{u}_i - \mathbf{u}_j)\| \\ &= \sqrt{\|\mathbf{a}_i - \mathbf{a}_j\|^2 + 2\varepsilon (\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j) + \varepsilon^2 \|\mathbf{u}_i - \mathbf{u}_j\|^2}. \end{aligned} \quad (6.44)$$



**Figure 6.8.** Tangent Line Approximation.

The difference between the new length and the original length, namely

$$e = \sqrt{\|\mathbf{a}_i - \mathbf{a}_j\|^2 + 2\varepsilon(\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j) + \varepsilon^2 \|\mathbf{u}_i - \mathbf{u}_j\|^2} - \|\mathbf{a}_i - \mathbf{a}_j\|, \quad (6.45)$$

is, by definition, the bar's *elongation*.

If the underlying dimension  $d$  is 2 or more, the elongation (6.45) is a *nonlinear* function of the displacement vectors  $\mathbf{u}_i, \mathbf{u}_j$ . Thus, an exact, geometrical treatment of structures in equilibrium requires dealing with complicated nonlinear systems of equations. In some situations, e.g., the design of robotic mechanisms, [57, 75], analysis of the nonlinear system is crucial, but this lies beyond the scope of this text. However, in many practical situations, the displacements are fairly small, so  $|\varepsilon| \ll 1$ . For example, when a building moves, the lengths of bars are in meters, but the displacements are, barring catastrophes, typically in centimeters if not millimeters. In such situations, we can replace the geometrically exact elongation by a much simpler linear approximation.

As you learned in calculus, the most basic linear approximation to a nonlinear function  $g(\varepsilon)$  near  $\varepsilon = 0$  is given by its tangent line or linear Taylor polynomial

$$g(\varepsilon) \approx g(0) + g'(0)\varepsilon, \quad |\varepsilon| \ll 1, \quad (6.46)$$

as sketched in [Figure 6.8](#). In the case of small displacements of a bar, the elongation (6.45) is a square root function of the particular form

$$g(\varepsilon) = \sqrt{a^2 + 2\varepsilon b + \varepsilon^2 c^2} - a,$$

where

$$a = \|\mathbf{a}_i - \mathbf{a}_j\|, \quad b = (\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j), \quad c = \|\mathbf{u}_i - \mathbf{u}_j\|,$$

are independent of  $\varepsilon$ . Since  $g(0) = 0$  and  $g'(0) = b/a$ , the linear approximation (6.46) is

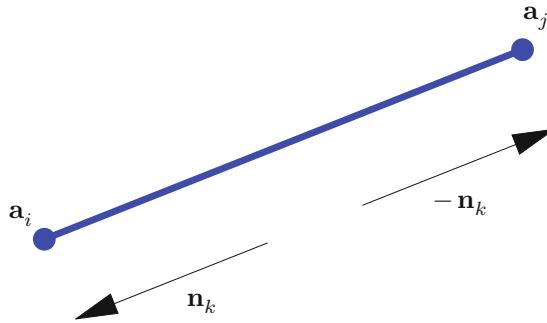
$$\sqrt{a^2 + 2\varepsilon b + \varepsilon^2 c^2} - a \approx \varepsilon \frac{b}{a} \quad \text{for} \quad |\varepsilon| \ll 1.$$

In this manner, we arrive at the linear approximation to the bar's elongation

$$e \approx \varepsilon \frac{(\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j)}{\|\mathbf{a}_i - \mathbf{a}_j\|} = \mathbf{n} \cdot (\varepsilon \mathbf{u}_i - \varepsilon \mathbf{u}_j), \quad \text{where} \quad \mathbf{n} = \frac{\mathbf{a}_i - \mathbf{a}_j}{\|\mathbf{a}_i - \mathbf{a}_j\|}$$

is the unit vector,  $\|\mathbf{n}\| = 1$ , that points in the direction of the bar from node  $j$  to node  $i$ .

The factor  $\varepsilon$  was merely a mathematical device used to derive the linear approximation. It can now be safely discarded, so that the displacement of the  $i^{\text{th}}$  node is now  $\mathbf{u}_i$  instead of  $\varepsilon \mathbf{u}_i$ , and we assume  $\|\mathbf{u}_i\|$  is small. If bar  $k$  connects node  $i$  to node  $j$ , then its (approximate)



**Figure 6.9.** Unit Vectors for a Bar.

elongation is equal to

$$e_k = \mathbf{n}_k \cdot (\mathbf{u}_i - \mathbf{u}_j) = \mathbf{n}_k \cdot \mathbf{u}_i - \mathbf{n}_k \cdot \mathbf{u}_j, \quad \text{where} \quad \mathbf{n}_k = \frac{\mathbf{a}_i - \mathbf{a}_j}{\|\mathbf{a}_i - \mathbf{a}_j\|}. \quad (6.47)$$

The elongation  $e_k$  is the sum of two terms: the first,  $\mathbf{n}_k \cdot \mathbf{u}_i$ , is the component of the displacement vector for node  $i$  in the direction of the unit vector  $\mathbf{n}_k$  that points along the bar *towards* node  $i$ , whereas the second,  $-\mathbf{n}_k \cdot \mathbf{u}_j$ , is the component of the displacement vector for node  $j$  in the direction of the unit vector  $-\mathbf{n}_k$  that points in the opposite direction along the bar *toward* node  $j$ ; see Figure 6.9. Their sum equals the total elongation of the bar.

We assemble all the linear equations (6.47) relating nodal displacements to bar elongations in matrix form

$$\mathbf{e} = A \mathbf{u}. \quad (6.48)$$

Here  $\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix} \in \mathbb{R}^m$  is the vector of elongations, while  $\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \in \mathbb{R}^{dn}$  is the vector

of displacements. Each  $\mathbf{u}_i \in \mathbb{R}^d$  is itself a column vector with  $d$  entries, and so  $\mathbf{u}$  has a total of  $dn$  entries. For example, in the planar case  $d = 2$ , we have  $\mathbf{u}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$ , since each node's displacement has both an  $x$  and  $y$  component, and so

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_n \\ y_n \end{pmatrix} \in \mathbb{R}^{2n}.$$

In three dimensions,  $d = 3$ , we have  $\mathbf{u}_i = (x_i, y_i, z_i)^T$ , and so each node will contribute three components to the displacement vector

$$\mathbf{u} = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, x_n, y_n, z_n)^T \in \mathbb{R}^{3n}.$$

The *incidence matrix*  $A$  connecting the displacements and elongations will be of size  $m \times (dn)$ . The  $k^{\text{th}}$  row of  $A$  will have (at most)  $2d$  nonzero entries. The entries in the  $d$

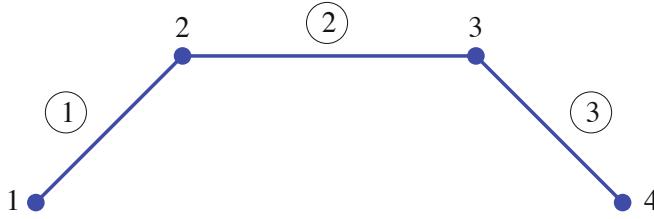


Figure 6.10. Three Bar Planar Structure.

slots corresponding to node  $i$  will be the components of the (transposed) unit bar vector  $\mathbf{n}_k^T$  pointing towards node  $i$ , as given in (6.47), while the entries in the  $d$  slots corresponding to node  $j$  will be the components of its negative  $-\mathbf{n}_k^T$ , which is the unit bar vector pointing towards node  $j$ . All other entries are 0. The general mathematical formulation is best appreciated by working through an explicit example.

**Example 6.5.** Consider the planar structure pictured in Figure 6.10. The four nodes are at positions

$$\mathbf{a}_1 = (0, 0)^T, \quad \mathbf{a}_2 = (1, 1)^T, \quad \mathbf{a}_3 = (3, 1)^T, \quad \mathbf{a}_4 = (4, 0)^T,$$

so the two side bars are at  $45^\circ$  angles and the center bar is horizontal. Implementing our construction, the associated incidence matrix is

$$A = \left( \begin{array}{cc|cc|cc|cc} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{array} \right). \quad (6.49)$$

The three rows of  $A$  refer to the three bars in our structure. The columns come in pairs, as indicated by the vertical lines in the matrix: the first two columns refer to the  $x$  and  $y$  displacements of the first node; the third and fourth columns refer to the second node; and so on. The first two entries of the first row of  $A$  indicate the unit vector

$$\mathbf{n}_1 = \frac{\mathbf{a}_1 - \mathbf{a}_2}{\|\mathbf{a}_1 - \mathbf{a}_2\|} = \left( -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)^T$$

that points along the first bar towards the first node, while the third and fourth entries have the opposite signs, and form the unit vector

$$-\mathbf{n}_1 = \frac{\mathbf{a}_2 - \mathbf{a}_1}{\|\mathbf{a}_2 - \mathbf{a}_1\|} = \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T$$

along the same bar that points in the opposite direction — towards the second node. The remaining entries are zero because the first bar connects only the first two nodes. Similarly, the unit vector along the second bar pointing towards node 2 is

$$\mathbf{n}_2 = \frac{\mathbf{a}_2 - \mathbf{a}_3}{\|\mathbf{a}_2 - \mathbf{a}_3\|} = (-1, 0)^T,$$

and this gives the third and fourth entries of the second row of  $A$ ; the fifth and sixth entries are their negatives, corresponding to the unit vector  $-\mathbf{n}_2$  pointing towards node 3. The last row is constructed from the unit vectors along bar #3 in the same fashion.

**Remark.** Interestingly, the incidence matrix for a structure depends only on the directions of the bars and not their lengths. This is analogous to the fact that the incidence matrix

for an electrical network depends only on the connectivity properties of the wires and not on their overall lengths. One can regard the incidence matrix for a structure as a kind of  $d$ -dimensional generalization of the incidence matrix for a directed graph.

The next phase of our procedure is to introduce the constitutive relations for the bars that determine their internal forces or stresses. As we remarked at the beginning of the section, each bar is viewed as a hard spring, subject to a linear Hooke's law equation

$$y_k = c_k e_k \quad (6.50)$$

that relates its elongation  $e_k$  to its internal force  $y_k$ . The bar stiffness  $c_k > 0$  is a positive scalar, and so  $y_k > 0$  if the bar is in tension, while  $y_k < 0$  if the bar is compressed. We write (6.50) in matrix form

$$\mathbf{y} = C \mathbf{e}, \quad (6.51)$$

where  $C = \text{diag}(c_1, \dots, c_m) > 0$  is a diagonal, positive definite matrix.

Finally, we need to balance the forces at each node in order to achieve equilibrium. If bar  $k$  terminates at node  $i$ , then it exerts a force  $-y_k \mathbf{n}_k$  on the node, where  $\mathbf{n}_k$  is the unit vector pointing towards the node in the direction of the bar, as in (6.47). The minus sign comes from physics: if the bar is under tension, so  $y_k > 0$ , then it is trying to contract back to its unstressed state, and so will pull the node towards it — in the opposite direction to  $\mathbf{n}_k$  — while a bar in compression will push the node away. In addition, we may have an externally applied force vector, denoted by  $\mathbf{f}_i$ , on node  $i$ , which might be some combination of gravity, weights, mechanical forces, and so on. (In this admittedly simplified model, external forces act only on the nodes and not directly on the bars.) Force balance at equilibrium requires that all the nodal forces, external and internal, cancel; thus,

$$\mathbf{f}_i + \sum_k (-y_k \mathbf{n}_k) = \mathbf{0}, \quad \text{or} \quad \sum_k y_k \mathbf{n}_k = \mathbf{f}_i,$$

where the sum is over all the bars that are attached to node  $i$ . The matrix form of the force balance equations is (and this should no longer come as a surprise)

$$\mathbf{f} = A^T \mathbf{y}, \quad (6.52)$$

where  $A^T$  is the transpose of the incidence matrix, and  $\mathbf{f} = \begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_n \end{pmatrix} \in \mathbb{R}^{dn}$  is the vector containing all external forces on the nodes. Putting everything together, (6.48, 51, 52),

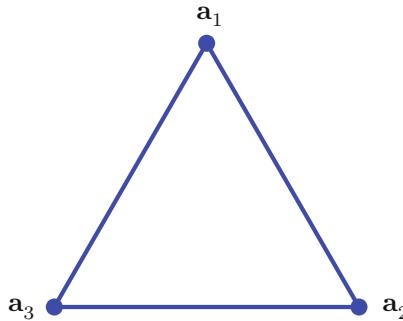
$$\mathbf{e} = A \mathbf{u}, \quad \mathbf{y} = C \mathbf{e}, \quad \mathbf{f} = A^T \mathbf{y},$$

we are once again led to a by now familiar linear system of equations:

$$K \mathbf{u} = \mathbf{f}, \quad \text{where} \quad K = A^T C A \quad (6.53)$$

is the stiffness matrix for our structure.

The stiffness matrix  $K$  is a positive (semi-)definite Gram matrix (3.64) associated with the weighted inner product on the space of elongations prescribed by the diagonal matrix  $C$ . As we know,  $K$  will be positive definite if and only if the kernel of the incidence matrix is trivial:  $\ker A = \{\mathbf{0}\}$ . However, the preceding example does not enjoy this property, because we have not tied down (or “grounded”) our structure. In essence, we are considering a structure floating in outer space, which is free to move around in any direction. Each rigid



**Figure 6.11.** A Triangular Structure.

motion<sup>†</sup> of the structure will correspond to an element of the kernel of its incidence matrix, and thereby preclude positive definiteness of its stiffness matrix.

**Example 6.6.** Consider a planar space station in the shape of a unit equilateral triangle, as in Figure 6.11. Placing the nodes at positions

$$\mathbf{a}_1 = \left( \frac{1}{2}, \frac{\sqrt{3}}{2} \right)^T, \quad \mathbf{a}_2 = (1, 0)^T, \quad \mathbf{a}_3 = (0, 0)^T,$$

we use the preceding algorithm to construct the incidence matrix

$$A = \left( \begin{array}{cc|cc|cc} -\frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & 0 \\ \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 & 1 & 0 & -1 & 0 \end{array} \right),$$

whose rows are indexed by the bars, and whose columns are indexed in pairs by the three nodes. The kernel of  $A$  is three-dimensional, with basis

$$\mathbf{z}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{z}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{z}_3 = \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \quad (6.54)$$

We claim that these three displacement vectors represent three different planar rigid motions: the first two correspond to translations, and the third to a rotation.

The translations are easy to discern. Translating the space station in a horizontal direction means that we move all three nodes the same amount, and so the displacements are  $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}_3 = \mathbf{a}$  for some vector  $\mathbf{a}$ . In particular, a rigid unit horizontal translation has  $\mathbf{a} = \mathbf{e}_1 = (1, 0)^T$ , and corresponds to the first kernel basis vector. Similarly, a unit vertical translation of all three nodes corresponds to  $\mathbf{a} = \mathbf{e}_2 = (0, 1)^T$ , and corresponds to the second kernel basis vector. Every other translation is a linear combination of these two. Translations do not alter the lengths of any of the bars, and so do not induce any stress in the structure.

<sup>†</sup> See Section 7.2 for an extended discussion of rigid motions.

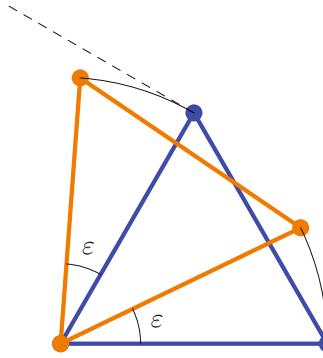


Figure 6.12. Rotating a Space Station.

The rotations are a little more subtle, owing to the linear approximation that we used to compute the elongations. Referring to Figure 6.12, we see that rotating the space station through a small angle  $\varepsilon$  around the node  $\mathbf{a}_3 = (0, 0)^T$  will move the other two nodes to positions

$$\mathbf{b}_1 = \begin{pmatrix} \frac{1}{2} \cos \varepsilon - \frac{\sqrt{3}}{2} \sin \varepsilon \\ \frac{1}{2} \sin \varepsilon + \frac{\sqrt{3}}{2} \cos \varepsilon \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} \cos \varepsilon \\ \sin \varepsilon \end{pmatrix}, \quad \mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (6.55)$$

However, the corresponding displacements

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{b}_1 - \mathbf{a}_1 = \begin{pmatrix} \frac{1}{2}(\cos \varepsilon - 1) - \frac{\sqrt{3}}{2} \sin \varepsilon \\ \frac{1}{2} \sin \varepsilon + \frac{\sqrt{3}}{2}(\cos \varepsilon - 1) \end{pmatrix}, \\ \mathbf{u}_2 &= \mathbf{b}_2 - \mathbf{a}_2 = \begin{pmatrix} \cos \varepsilon - 1 \\ \sin \varepsilon \end{pmatrix}, \quad \mathbf{u}_3 = \mathbf{b}_3 - \mathbf{a}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \end{aligned} \quad (6.56)$$

do *not* combine into a vector that belongs to  $\ker A$ . The problem is that, under a rotation, the nodes move along circles, while the kernel displacements  $\mathbf{u} = \varepsilon \mathbf{z} \in \ker A$  correspond to straight line motion! In order to maintain consistency, we must adopt a similar linear approximation of the nonlinear circular motion of the nodes. Thus, we replace the nonlinear displacements  $\mathbf{u}_j(\varepsilon)$  in (6.56) by their linear tangent approximations<sup>†</sup>  $\varepsilon \mathbf{u}'_j(0)$ , so

$$\mathbf{u}_1 \approx \varepsilon \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix}, \quad \mathbf{u}_2 \approx \varepsilon \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The resulting displacements *do* combine to produce the displacement vector

$$\mathbf{u} = \varepsilon \begin{pmatrix} -\frac{\sqrt{3}}{2}, \frac{1}{2}, 0, 1, 0, 0 \end{pmatrix}^T = \varepsilon \mathbf{z}_3$$

that moves the space station in the direction of the third kernel basis vector. Thus, as claimed,  $\mathbf{z}_3$  represents the linear approximation to a rigid rotation around the first node.

Remarkably, the rotations around the other two nodes, although distinct nonlinear motions, can be linearly approximated by particular combinations of the three kernel basis

---

<sup>†</sup> Note that  $\mathbf{u}_j(0) = \mathbf{0}$ .

elements  $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ , and so already appear in our description of  $\ker A$ . For example, the displacement vector

$$\mathbf{u} = \varepsilon \left( \frac{\sqrt{3}}{2} \mathbf{z}_1 + \frac{1}{2} \mathbf{z}_2 - \mathbf{z}_3 \right) = \varepsilon \left( 0, 0, \frac{\sqrt{3}}{2}, -\frac{1}{2}, \frac{\sqrt{3}}{2}, \frac{1}{2} \right)^T \quad (6.57)$$

represents the linear approximation to a rigid rotation around the first node. We conclude that the three-dimensional kernel of the incidence matrix represents the sum total of all possible rigid motions of the space station, or, more correctly, their linear approximations.

Which types of forces will maintain the space station in equilibrium? This will happen if and only if we can solve the force balance equations  $A^T \mathbf{y} = \mathbf{f}$  for the internal forces  $\mathbf{y}$ . The Fredholm Alternative Theorem 4.46 implies that this system has a solution if and only if  $\mathbf{f}$  is orthogonal to  $\text{coker } A^T = \ker A$ . Therefore,  $\mathbf{f} = (f_1, g_1, f_2, g_2, f_3, g_3)^T$  must be orthogonal to the kernel basis vectors (6.54), and so must satisfy the three linear constraints

$$\begin{aligned} \mathbf{z}_1 \cdot \mathbf{f} &= f_1 + f_2 + f_3 = 0, \\ \mathbf{z}_2 \cdot \mathbf{f} &= g_1 + g_2 + g_3 = 0, \\ \mathbf{z}_3 \cdot \mathbf{f} &= -\frac{\sqrt{3}}{2} f_1 + \frac{1}{2} g_1 + g_2 = 0. \end{aligned} \quad (6.58)$$

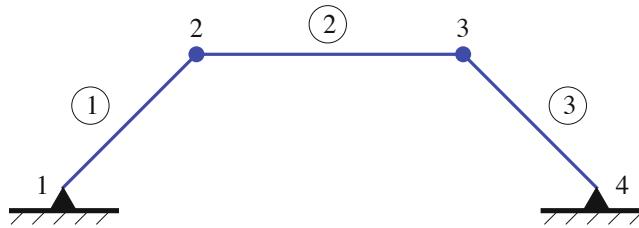
The first constraint requires that there be no net horizontal force on the space station. The second requires no net vertical force. The last constraint requires that the *moment* of the forces around the third node vanishes. The vanishing of the force moments around each of the other two nodes is a consequence of these three conditions, since the associated kernel vectors can be expressed as linear combinations of the three basis elements. The corresponding physical requirements are clear. If there is a net horizontal or vertical force, the space station will rigidly translate in that direction; if there is a non-zero force moment, the station will rigidly rotate. In any event, unless the force balance constraints (6.58) are satisfied, the space station cannot remain in equilibrium. A freely floating space station is an unstable structure that can easily be set into motion with a tiny external force.

Since there are three independent rigid motions, we must impose three constraints on the structure in order to fully stabilize it under general external forcing. “Grounding” one of the nodes, i.e., preventing it from moving by attaching it to a fixed support, will serve to eliminate the two translational instabilities. For example, setting  $\mathbf{u}_3 = \mathbf{0}$  has the effect of fixing the third node of the space station to a support. With this specification, we can eliminate the variables associated with that node, and thereby delete the corresponding columns of the incidence matrix — leaving the *reduced incidence matrix*

$$A^* = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The kernel of  $A^*$  is one-dimensional, spanned by the single vector  $\mathbf{z}_3^* = \left( \frac{\sqrt{3}}{2}, \frac{1}{2}, 0, 1 \right)^T$ , which corresponds to (the linear approximation of) the rotations around the fixed node. To prevent the structure from rotating, we can also fix the second node, by further requiring  $\mathbf{u}_2 = \mathbf{0}$ . This serves to also eliminate the third and fourth columns of the original incidence matrix. The resulting “doubly reduced” incidence matrix

$$A^{**} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 \end{pmatrix}$$



**Figure 6.13.** Three Bar Structure with Fixed Supports.

has trivial kernel:  $\ker A^{**} = \{\mathbf{0}\}$ . Therefore, the corresponding reduced stiffness matrix

$$K^{**} = (A^{**})^T A^{**} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{pmatrix}$$

is positive definite. A planar triangle with two fixed nodes is a stable structure, which can now support an arbitrary external forcing on the remaining free node. (Forces on the fixed nodes have no effect, since they are no longer allowed to move.)

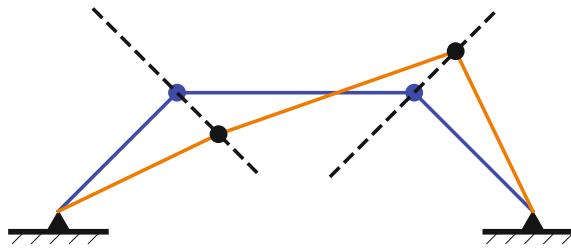
In general, a planar structure without any fixed nodes will have at least a three-dimensional kernel, corresponding to the rigid motions of translations and (linear approximations to) rotations. To stabilize the structure, one must fix two (non-coincident) nodes. A three-dimensional structure that is not tied to any fixed supports will admit 6 independent rigid motions in its kernel. Three of these correspond to rigid translations in the three coordinate directions, while the other three correspond to linear approximations to the rigid rotations around the three coordinate axes. To eliminate the rigid motion instabilities of the structure, we need to fix three non-collinear nodes. Indeed, fixing one node will eliminate translations; fixing two nodes will still leave the rotations around the axis through the fixed nodes. Details can be found in the exercises.

Even after a sufficient number of nodes have been attached to fixed supports so as to eliminate all possible rigid motions, there may still remain nonzero vectors in the kernel of the reduced incidence matrix of the structure. These indicate additional instabilities that allow the shape of the structure to deform without any applied force. Such non-rigid motions are known as *mechanisms* of the structure. Since a mechanism moves the nodes without elongating any of the bars, it does not induce any internal forces. A structure that admits a mechanism is unstable — even tiny external forces may provoke a large motion.

**Example 6.7.** Consider the three-bar structure of Example 6.5, but now with its two ends attached to supports, as pictured in [Figure 6.13](#). Since we are fixing nodes 1 and 4, we set  $\mathbf{u}_1 = \mathbf{u}_4 = \mathbf{0}$ . Hence, we should remove the first and last column pairs from the incidence matrix (6.49), leading to the reduced incidence matrix

$$A^* = \left( \begin{array}{cc|cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array} \right).$$

The structure no longer admits any rigid motions. However, the kernel of  $A^*$  is one-dimensional, spanned by reduced displacement vector  $\mathbf{z}^* = (1, -1, 1, 1)^T$ , which corresponds to the unstable mechanism that displaces the second node in the direction  $\mathbf{u}_2 =$



**Figure 6.14.** Unstable Mechanism of the Three Bar Structure.

$(1, -1)^T$  and the third node in the direction  $\mathbf{u}_3 = (1, 1)^T$ . Geometrically, then,  $\mathbf{z}^*$  represents the displacement whereby node 2 moves down and to the right at a  $45^\circ$  angle, while node 3 moves simultaneously up and to the right at a  $45^\circ$  angle; the result of the mechanism is sketched in [Figure 6.14](#). This mechanism does not alter the lengths of the three bars (at least in our linear approximation regime) and so requires no net force to be set into motion.

As with the rigid motions of the space station, an external forcing vector  $\mathbf{f}^*$  will maintain equilibrium only when it lies in the coimage of  $A^*$ , and hence, by the Fredholm Alternative, must be orthogonal to all the mechanisms in  $\ker A^*$ . Thus, the nodal forces  $\mathbf{f}_2 = (f_2, g_2)^T$  and  $\mathbf{f}_3 = (f_3, g_3)^T$  must satisfy the balance law

$$\mathbf{z}^* \cdot \mathbf{f}^* = f_2 - g_2 + f_3 + g_3 = 0.$$

If this fails, the equilibrium equation has no solution, and the structure will be set into motion. For example, a uniform horizontal force  $f_2 = f_3 = 1$ ,  $g_2 = g_3 = 0$ , will induce the mechanism, whereas a uniform vertical force,  $f_2 = f_3 = 0$ ,  $g_2 = g_3 = 1$ , will maintain equilibrium. In the latter case, the equilibrium equations

$$K^* \mathbf{u}^* = \mathbf{f}^*, \quad \text{where} \quad K^* = (A^*)^T A^* = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} & -1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ -1 & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

have an indeterminate solution

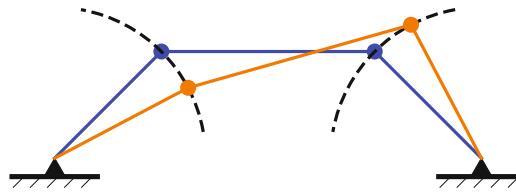
$$\mathbf{u}^* = (-3, 5, -2, 0)^T + t (1, -1, 1, 1)^T,$$

since we can add in any element of  $\ker K^* = \ker A^*$ . In other words, the equilibrium position is not unique, since the structure can still be displaced in the direction of the unstable mechanism while maintaining the overall force balance. On the other hand, the elongations and internal forces

$$\mathbf{y} = \mathbf{e} = A^* \mathbf{u}^* = (\sqrt{2}, 1, \sqrt{2})^T,$$

are well defined, indicating that, under our stabilizing uniform vertical (upwards) force, all three bars are elongated, with the two diagonals experiencing 41.4% more elongation than the horizontal bar.

**Remark.** Just like the rigid rotations, the mechanisms described here are linear approximations to the actual nonlinear motions. In a physical structure, the vertices will move



**Figure 6.15.** Nonlinear Mechanism of the Three Bar Structure.

along curves whose tangents at the initial configuration are the directions indicated by the mechanism vector. In the linear approximation illustrated in Figure 6.14, the lengths of the bars will change slightly. In the true nonlinear mechanism, illustrated in Figure 6.15, the nodes must move along circles so as to rigidly preserve the lengths of all three bars. In certain cases, a structure can admit a linear mechanism, but one that cannot be physically realized due to the nonlinear constraints imposed by the geometrical configurations of the bars. Nevertheless, such a structure is at best borderline stable, and should not be used in any real-world constructions.

We can always stabilize a structure by first fixing nodes to eliminate rigid motions, and then adding in a sufficient number of extra bars to prevent mechanisms. In the preceding example, suppose we attach an additional bar connecting nodes 2 and 4, leading to the reinforced structure in Figure 6.16. The revised incidence matrix is

$$A = \left( \begin{array}{cc|cc|cc|cc} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 & \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{array} \right),$$

and is obtained from (6.49) by appending another row representing the added bar. When nodes 1 and 4 are fixed, the reduced incidence matrix

$$A^* = \left( \begin{array}{ccccc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 \end{array} \right)$$

has trivial kernel,  $\ker A^* = \{\mathbf{0}\}$ , and hence the reinforced structure is stable. It admits no mechanisms, and can support any configuration of forces (within reason — mathematically the structure will support an arbitrarily large external force, but very large forces will take us outside the linear regime described by the model, and the structure may be crushed).

This particular case is *statically determinate* owing to the fact that the incidence matrix is square and nonsingular, which implies that one can solve the force balance equations (6.52) directly for the internal forces. For instance, a uniform downwards vertical force  $f_2 = f_3 = 0$ ,  $g_2 = g_3 = -1$ , e.g., gravity, will produce the internal forces

$$y_1 = -\sqrt{2}, \quad y_2 = -1, \quad y_3 = -\sqrt{2}, \quad y_4 = 0,$$

indicating that bars 1, 2 and 3 are compressed, while, interestingly, the reinforcing bar 4 remains unchanged in length and hence experiences no internal force. Assuming that the

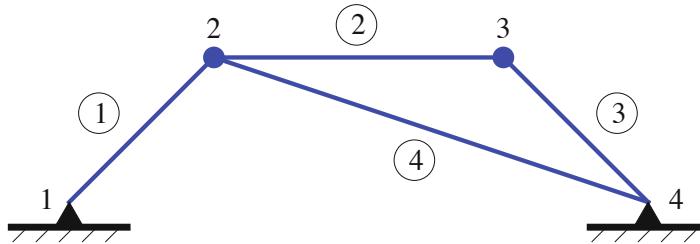


Figure 6.16. Reinforced Planar Structure.

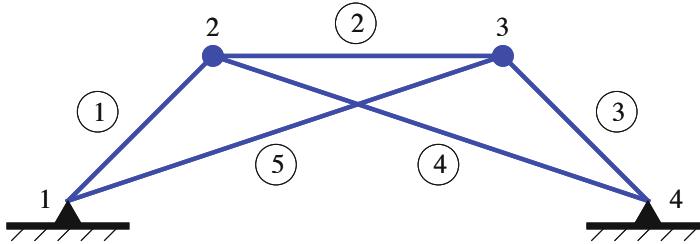


Figure 6.17. Doubly Reinforced Planar Structure.

bars are all of the same material, and taking the elastic constant to be 1, so  $C = I$ , then the reduced stiffness matrix is

$$K^* = (A^*)^T A^* = \begin{pmatrix} \frac{12}{5} & \frac{1}{5} & -1 & 0 \\ \frac{1}{5} & \frac{3}{5} & 0 & 0 \\ -1 & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The solution to the reduced equilibrium equations is

$$\mathbf{u}^* = \left( -\frac{1}{2}, -\frac{3}{2}, -\frac{3}{2}, -\frac{7}{2} \right)^T, \quad \text{so} \quad \mathbf{u}_2 = \left( -\frac{1}{2}, -\frac{3}{2} \right)^T, \quad \mathbf{u}_3 = \left( -\frac{3}{2}, -\frac{7}{2} \right)^T,$$

give the displacements of the two nodes under the applied force. Both are moving down and to the left, with node 3 moving relatively farther owing to its lack of reinforcement.

Suppose we reinforce the structure yet further by adding in a bar connecting nodes 1 and 3, as in Figure 6.17. The resulting reduced incidence matrix

$$A^* = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 \\ 0 & 0 & \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix}$$

again has trivial kernel,  $\ker A^* = \{\mathbf{0}\}$ , and hence the structure is stable. Indeed, adding extra bars to a stable structure cannot cause it to lose stability. (In the language of linear algebra, appending additional rows to a matrix cannot increase the size of its kernel, cf. Exercise 2.5.10.) Since the incidence matrix is rectangular, the structure is now *statically indeterminate*, and we cannot determine the internal forces without first solving the full

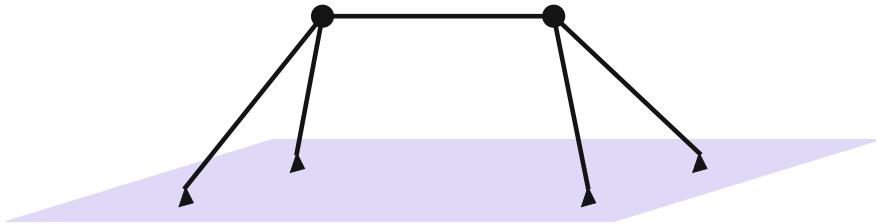


Figure 6.18. A Swing Set.

equilibrium equations (6.53) for the displacements. The stiffness matrix is

$$K^* = (A^*)^T A^* = \begin{pmatrix} \frac{12}{5} & \frac{1}{5} & -1 & 0 \\ \frac{1}{5} & \frac{3}{5} & 0 & 0 \\ -1 & 0 & \frac{12}{5} & -\frac{1}{5} \\ 0 & 0 & -\frac{1}{5} & \frac{3}{5} \end{pmatrix}.$$

Under the same uniform vertical force, the displacement  $\mathbf{u}^* = (\frac{1}{10}, -\frac{17}{10}, -\frac{1}{10}, -\frac{17}{10})^T$  indicates that the free nodes now move symmetrically down and towards the center of the structure. The internal forces on the bars are

$$y_1 = -\frac{4}{5}\sqrt{2}, \quad y_2 = -\frac{1}{5}, \quad y_3 = -\frac{4}{5}\sqrt{2}, \quad y_4 = -\sqrt{\frac{2}{5}}, \quad y_5 = -\sqrt{\frac{2}{5}}.$$

All five bars are now experiencing compression, with the two outside bars being the most stressed. This relatively simple computation should already indicate to the practicing construction engineer which of the bars in the structure are more likely to collapse under an applied external force.

Summarizing our discussion, we have established the following fundamental result characterizing the stability and equilibrium of structures.

**Theorem 6.8.** A structure is stable, and so will maintain its equilibrium under arbitrary external forcing, if and only if its reduced incidence matrix  $A^*$  has linearly independent columns, or, equivalently,  $\ker A^* = \{\mathbf{0}\}$ . More generally, an external force  $\mathbf{f}^*$  on a structure will maintain equilibrium if and only if  $\mathbf{f}^* \in \text{coimg } A^* = (\ker A^*)^\perp$ , which requires that the external force be orthogonal to all rigid motions and all mechanisms admitted by the structure.

**Example 6.9.** A three-dimensional swing set is to be constructed, consisting of two diagonal supports at each end joined by a horizontal cross bar. Is this configuration stable, i.e., can a child swing on it without it collapsing? The movable joints are at positions

$$\mathbf{a}_1 = (1, 1, 3)^T, \quad \mathbf{a}_2 = (4, 1, 3)^T,$$

while the four fixed supports are at

$$\mathbf{a}_3 = (0, 0, 0)^T, \quad \mathbf{a}_4 = (0, 2, 0)^T, \quad \mathbf{a}_5 = (5, 0, 0)^T, \quad \mathbf{a}_6 = (5, 2, 0)^T.$$

The reduced incidence matrix for the structure is calculated in the usual manner:

$$A^* = \left( \begin{array}{ccc|ccc} \frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \end{array} \right).$$

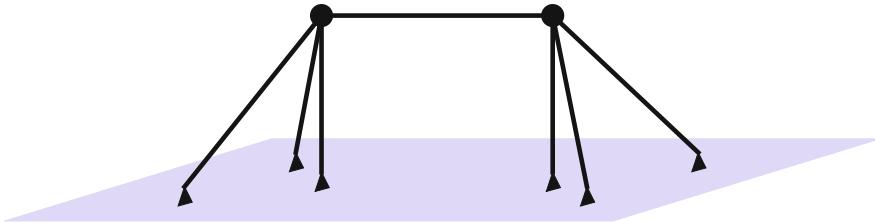


Figure 6.19. Reinforced Swing Set.

For instance, the first three entries contained in the first row refer to the unit vector  $\mathbf{n}_1 = \frac{\mathbf{a}_1 - \mathbf{a}_3}{\|\mathbf{a}_1 - \mathbf{a}_3\|}$  in the direction of the bar going from  $\mathbf{a}_3$  to  $\mathbf{a}_1$ . Suppose the five bars have the same stiffness  $c_1 = \dots = c_5 = 1$ , so the reduced stiffness matrix for the structure is

$$K^* = (A^*)^T A^* = \begin{pmatrix} \frac{13}{11} & 0 & \frac{6}{11} & -1 & 0 & 0 \\ 0 & \frac{2}{11} & 0 & 0 & 0 & 0 \\ \frac{6}{11} & 0 & \frac{18}{11} & 0 & 0 & 0 \\ -1 & 0 & 0 & \frac{13}{11} & 0 & -\frac{6}{11} \\ 0 & 0 & 0 & 0 & \frac{2}{11} & 0 \\ 0 & 0 & 0 & -\frac{6}{11} & 0 & \frac{18}{11} \end{pmatrix}.$$

Solving  $A^* \mathbf{z}^* = \mathbf{0}$ , we find  $\ker A^* = \ker K^*$  is one-dimensional, spanned by

$$\mathbf{z}^* = (3, 0, -1, 3, 0, 1)^T.$$

This indicates a mechanism that can cause the swing set to collapse: the first node moves down and to the right, while the second node moves up and to the right, the horizontal motion being three times as large as the vertical. The swing set can only support forces  $\mathbf{f}_1 = (f_1, g_1, h_1)^T$ ,  $\mathbf{f}_2 = (f_2, g_2, h_2)^T$  on the free nodes whose combined force vector  $\mathbf{f}^*$  is orthogonal to the mechanism vector  $\mathbf{z}^*$ , and so

$$3(f_1 + f_2) - h_1 + h_2 = 0.$$

Otherwise, a reinforcing bar, say from node 1 to node 6 (although this will interfere with the swinging!) or another bar connecting one of the nodes to a new ground support, will be required to completely stabilize the swing.

For a uniform downwards unit vertical force,  $\mathbf{f} = (0, 0, -1, 0, 0, -1)^T$ , a particular solution to (6.11) is  $\mathbf{u}^* = (\frac{13}{6}, 0, -\frac{4}{3}, \frac{11}{6}, 0, 0)^T$  and the general solution  $\mathbf{u} = \mathbf{u}^* + t \mathbf{z}^*$  is obtained by adding in an arbitrary element of the kernel. The resulting forces/elongations are uniquely determined,

$$\mathbf{y} = \mathbf{e} = A^* \mathbf{u} = A^* \mathbf{u}^* = \left( -\frac{\sqrt{11}}{6}, -\frac{\sqrt{11}}{6}, -\frac{1}{3}, -\frac{\sqrt{11}}{6}, -\frac{\sqrt{11}}{6} \right)^T,$$

so that every bar is compressed, the middle one experiencing slightly more than half the stress of the outer supports.

If we add in two vertical supports at the nodes, as in Figure 6.19, then the corresponding

reduced incidence matrix

$$A^* = \begin{pmatrix} \frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

has trivial kernel, indicating stabilization of the structure. The reduced stiffness matrix

$$K^* = \begin{pmatrix} \frac{13}{11} & 0 & \frac{6}{11} & -1 & 0 & 0 \\ 0 & \frac{2}{11} & 0 & 0 & 0 & 0 \\ \frac{6}{11} & 0 & \frac{29}{11} & 0 & 0 & 0 \\ -1 & 0 & 0 & \frac{13}{11} & 0 & -\frac{6}{11} \\ 0 & 0 & 0 & 0 & \frac{2}{11} & 0 \\ 0 & 0 & 0 & -\frac{6}{11} & 0 & \frac{29}{11} \end{pmatrix}$$

is now only slightly different, but this is enough to make it positive definite,  $K^* > 0$ , and so allow arbitrary external forcing without collapse. Under the same uniform vertical force, the internal forces are

$$\mathbf{y} = \mathbf{e} = A^* \mathbf{u} = \left( -\frac{\sqrt{11}}{10}, -\frac{\sqrt{11}}{10}, -\frac{1}{5}, -\frac{\sqrt{11}}{10}, -\frac{\sqrt{11}}{10}, -\frac{2}{5}, -\frac{2}{5} \right)^T.$$

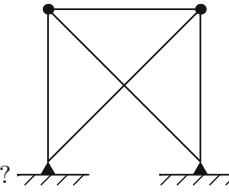
Note the overall reductions in stress in the original bars; the two reinforcing vertical bars are now experiencing the largest compression.

Further developments in the mathematical analysis of structures can be found in the references [33, 79].

## Exercises

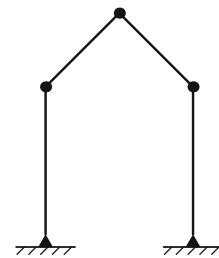
- 6.3.1. If a bar in a structure compresses 2 cm under a force of 5 newtons applied to a node, how far will it compress under a force of 20 newtons applied at the same node?
- 6.3.2. An individual bar in a structure experiences a stress of 3 under a unit horizontal force applied to all the nodes and a stress of  $-2$  under a unit vertical force applied to all nodes. What combinations of horizontal and vertical forces will make the bar stress-free?
- 6.3.3. (a) For the reinforced structure illustrated in [Figure 6.16](#), determine the displacements of the nodes and the stresses in the bars under a uniform horizontal force, and interpret physically. (b) Answer the same question for the doubly reinforced structure in [Figure 6.17](#).
- 6.3.4. Discuss the effect of a uniform horizontal force in the direction of the horizontal bar on the swing set and its reinforced version in Example 6.9.

- ◇ 6.3.5. All the bars in the illustrated square planar structure have unit stiffness. (a) Write down the reduced incidence matrix  $A$ . (b) Write down the equilibrium equations for the structure when subjected to external forces at the free nodes. (c) Is the structure stable? statically determinate? Explain in detail. (d) Find a set of external forces with the property that the upper left node moves horizontally, while the upper right node stays in place. Which bar is under the most stress?



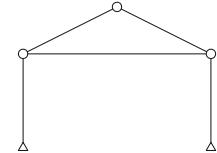
- ◇ 6.3.6. In the square structure of Exercise 6.3.5, the diagonal struts simply cross each other. We could also try joining them at an additional central node. Compare the stresses in the two structures under a uniform horizontal and a uniform vertical force at the two upper nodes, and discuss what you observe.

- 6.3.7. (a) Write down the reduced incidence matrix  $A^*$  for the pictured structure with 4 bars and 2 fixed supports. The width and the height of the vertical sides are each 1 unit, while the top node is 1.5 units above the base. (b) Predict the number of independent solutions to  $A^* \mathbf{u} = \mathbf{0}$ , and then solve to describe them both numerically and geometrically. (c) What condition(s) must be imposed on the external forces to maintain equilibrium in the structure? (d) Add in just enough additional bars so that the resulting reinforced structure has only the trivial solution to  $A^* \mathbf{u} = \mathbf{0}$ . Is your reinforced structure stable?

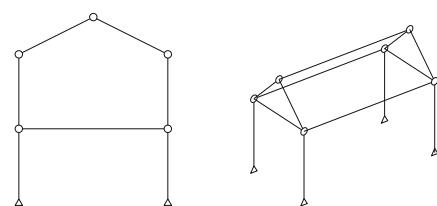


- ◇ 6.3.8. Consider the two-dimensional “house” constructed out of bars, as in the accompanying picture. The bottom nodes are fixed. The width of the house is 3 units, the height of the vertical sides 1 unit, and the peak is 1.5 units above the base.

- (a) Determine the reduced incidence matrix  $A$  for this structure.  
 (b) How many distinct modes of instability are there? Describe them geometrically, and indicate whether they are mechanisms or rigid motions.  
 (c) Suppose we apply a combination of forces to each non-fixed node in the structure. Determine conditions such that the structure can support the forces. Write down an explicit nonzero set of external forces that satisfy these conditions, and compute the corresponding elongations of the individual bars. Which bar is under the most stress?  
 (d) Add in a *minimal* number of bars so that the resulting structure can support any force. Before starting, decide, from general principles, how many bars you need to add.  
 (e) With your new stable configuration, use the same force as before, and recompute the forces on the individual bars. Which bar now has the most stress? How much have you reduced the maximal stress in your reinforced building?



- ♣ 6.3.9. Answer Exercise 6.3.8 for the illustrated two- and three-dimensional houses. In the two-dimensional case, the width and total height of the vertical bars is 2 units, and the peak is an additional .5 unit higher. In the three-dimensional house, the width and vertical heights are equal to 1 unit, the length is 3 units, while the peaks are 1.5 units above the base.



- ◇ 6.3.10. Consider a structure consisting of three bars joined in a vertical line hanging from a top support. (a) Write down the equilibrium equations for this system when only forces and displacements in the vertical direction are allowed, i.e., a one-dimensional structure. Is the problem statically determinate, statically indeterminate, or unstable? If the latter, describe all possible mechanisms and the constraints on the forces required to maintain equilibrium.  
 (b) Answer part (a) when the structure is two-dimensional, i.e., is allowed to move in a plane. (c) Answer the same question for the fully three-dimensional version.

- ♣ 6.3.11. A space station is built in the shape of a three-dimensional *simplex* whose nodes are at the positions  $\mathbf{0}, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \in \mathbb{R}^3$ , and each pair of nodes is connected by a bar. (a) Sketch the space station and find its incidence matrix  $A$ . (b) Show that  $\ker A$  is six-dimensional, and find a basis. (c) Explain which three basis vectors correspond to rigid translations. (d) Find three basis vectors that correspond to linear approximations to rotations around the three coordinate axes. (e) Suppose the bars all have unit stiffness. Compute the full stiffness matrix for the space station. (f) What constraints on external forces at the four nodes are required to maintain equilibrium? Can you interpret them physically? (g) How many nodes do you need to fix to stabilize the structure? (h) Suppose you fix the three nodes in the  $xy$ -plane. How much internal force does each bar experience under a unit vertical force on the upper vertex?
- ♣ 6.3.12. Suppose a space station is built in the shape of a regular tetrahedron with all sides of unit length. Answer all questions in Exercise 6.3.11.
- ♡ 6.3.13. A *mass-spring ring* consists of  $n$  masses connected in a circle by  $n$  identical springs, and the masses are allowed only to move in the angular direction. (a) Derive the equations of equilibrium. (b) Discuss stability, and characterize the external forces that will maintain equilibrium. (c) Find such a set of nonzero external forces in the case of a four-mass ring and solve the equilibrium equations. What does the nonuniqueness of the solution represent?
- 6.3.14. A structure in  $\mathbb{R}^3$  has  $n$  movable nodes, admits no rigid motions, and is statically determinate. (a) How many bars must it have? (b) Find an example with  $n = 3$ .
- ◊ 6.3.15. Prove that if we apply a unit force to node  $i$  in a structure and measure the displacement of node  $j$  in the direction of the force, then we obtain the same value if we apply the force to node  $j$  and measure the displacement at node  $i$  in the same direction. *Hint:* First, solve Exercise 6.1.6.
- 6.3.16. *True or false:* A structure in  $\mathbb{R}^3$  will admit no rigid motions if and only if at least 3 nodes are fixed.
- 6.3.17. Suppose all bars have unit stiffness. Explain why the internal forces in a structure form the solution of minimal Euclidean norm among all solutions to  $A^T \mathbf{y} = \mathbf{f}$ .
- ◊ 6.3.18. Let  $A$  be the reduced incidence matrix for a structure and  $C$  the diagonal bar stiffness matrix. Suppose  $\mathbf{f}$  is a set of external forces that maintain equilibrium of the structure. (a) Prove that  $\mathbf{f} = A^T C \mathbf{g}$  for some  $\mathbf{g}$ . (b) Prove that an allowable displacement  $\mathbf{u}$  is a least squares solution to the system  $A \mathbf{u} = \mathbf{g}$  with respect to the weighted norm  $\|\mathbf{v}\|^2 = \mathbf{v}^T C \mathbf{v}$ .
- ♡ 6.3.19. Suppose an *unstable* structure admits no rigid motions — only mechanisms. Let  $\mathbf{f}$  be an external force on the structure that maintains equilibrium. Suppose that you stabilize the structure by adding in the *minimal* number of reinforcing bars. Prove that the given force  $\mathbf{f}$  induces the *same* stresses in the original bars, while the reinforcing bars experience no stress. Are the displacements necessarily the same? Does the result continue to hold when more reinforcing bars are added to the structure? *Hint:* Use Exercise 6.3.18.
- ♡ 6.3.20. When a node is fixed to a *roller*, it is permitted to move only along a straight line — the direction of the roller. Consider the three-bar structure in Example 6.5. Suppose node 1 is fixed, but node 4 is attached to a roller that permits it to move only in the horizontal direction. (a) Construct the reduced incidence matrix and the equilibrium equations in this situation. You should have a system of 5 equations in 5 unknowns — the horizontal and vertical displacements of nodes 2 and 3 and the horizontal displacement of node 4. (b) Is your structure stable? If not, how many rigid motions and how many mechanisms does it permit?
- ♡ 6.3.21. Answer Exercise 6.3.20 when the roller at node 4 allows it to move in only the vertical direction.

♡ 6.3.22. Redo Exercises 6.3.20–21 for the reinforced structure in [Figure 6.16](#).

6.3.23. (a) Suppose that we fix one node in a planar structure and put a second node on a roller. Does the structure admit any rigid motions? (b) How many rollers are needed to prevent all rigid motions in a three-dimensional structure? Are there any restrictions on the directions of the rollers?

6.3.24. *True or false:* If a structure is statically indeterminate, then every non-zero applied force will result in (a) one or more nodes having a non-zero displacement; (b) one or more bars having a non-zero elongation.

6.3.25. *True or false:* If a structure constructed out of bars with identical stiffnesses is stable, then the same structure constructed out of bars with differing stiffnesses is also stable.

---



# Chapter 7

## Linearity

We began this book by learning how to systematically solve systems of linear algebraic equations. This “elementary” problem formed our launching pad for developing the fundamentals of linear algebra. In its initial form, matrices and vectors were the primary focus of our study, but the theory was developed in a sufficiently general and abstract form that it can be immediately used in many other useful situations — particularly infinite-dimensional function spaces. Indeed, applied mathematics deals, not just with algebraic equations, but also with differential equations, difference equations, integral equations, stochastic systems, differential delay equations, control systems, and many other types — only a few of which, unfortunately, can be adequately developed in this introductory text. It is now time to assemble what we have learned about linear algebraic systems and place the results in a suitably general framework that will lead to insight into the key principles that govern all linear systems arising in mathematics and its applications.

The most basic underlying object of linear systems theory is the vector space, and we have already seen that the elements of vector spaces can be vectors, or functions, or even vector-valued functions. The seminal ideas of span, linear independence, basis, and dimension are equally applicable and equally vital in more general contexts, particularly function spaces. Just as vectors in Euclidean space are prototypes for elements of general vector spaces, matrices are also prototypes for more general objects, known as *linear functions*. Linear functions are also known as linear maps or, when one is dealing with function spaces, linear operators, and include linear differential operators, linear integral operators, function evaluation, and many other basic operations. Linear operators on infinite-dimensional function spaces are the basic objects of quantum mechanics. Each quantum mechanical observable (mass, energy, momentum) is formulated as a linear operator on an infinite-dimensional Hilbert space — the space of wave functions or states of the system, [54]. It is remarkable that quantum mechanics is an entirely linear theory, whereas classical and relativistic mechanics are inherently nonlinear. The holy grail of modern physics — the unification of general relativity and quantum mechanics — is to resolve the apparent incompatibility of the microscopic linear and macroscopic nonlinear physical regimes.

In geometry, linear functions are interpreted as linear transformations of space (or space-time), and, as such, lie at the foundations of motion of bodies, such as satellites and planets; computer graphics and games; video, animation, and movies; and the mathematical formulation of symmetry. Many familiar geometrical transformations, including rotations, scalings and stretches, reflections, projections, shears, and screw motions, are linear. But including translational motions requires a slight extension of linearity, known as an affine transformation. The basic geometry of linear and affine transformations will be developed in Section 7.2.

Linear functions form the simplest class of functions on vector spaces, and must be thoroughly understood before any serious progress can be made in the vastly more complicated nonlinear world. Indeed, nonlinear functions are often approximated by linear functions, generalizing the calculus approximation of a scalar function by its tangent line. This linearization process is applied to nonlinear functions of several variables studied in

multivariable calculus, as well as the nonlinear systems arising in physics and mechanics, which can often be well approximated by linear differential equations.

A linear system is just an equation formed by a linear function. The most basic linear system is a system of linear algebraic equations. Linear systems theory includes linear differential equations, linear boundary value problems, linear integral equations, and so on, all in a common conceptual framework. The fundamental ideas of linear superposition and the relation between the solutions to inhomogeneous and homogeneous systems are universally applicable to all linear systems. You have no doubt encountered many of these concepts in your study of elementary ordinary differential equations. In this text, they have already appeared in our discussion of the solutions to linear algebraic systems. The final section introduces the notion of the adjoint of a linear map between inner product spaces, generalizing the transpose operation on matrices, the notion of a positive definite linear operator, and the characterization of the solution to such a linear system by a minimization principle. The full import of these fundamental concepts will appear in the context of linear boundary value problems and partial differential equations, [61].

## 7.1 Linear Functions

We begin our study of linear functions with the basic definition. For simplicity, we shall concentrate on real linear functions between real vector spaces. Extending the concepts and constructions to complex linear functions on complex vector spaces is not difficult, and will be dealt with in due course.

**Definition 7.1.** Let  $V$  and  $W$  be real vector spaces. A function  $L: V \rightarrow W$  is called *linear* if it obeys two basic rules:

$$L[\mathbf{v} + \mathbf{w}] = L[\mathbf{v}] + L[\mathbf{w}], \quad L[c\mathbf{v}] = cL[\mathbf{v}], \quad (7.1)$$

for all  $\mathbf{v}, \mathbf{w} \in V$  and all scalars  $c$ . We will call  $V$  the *domain* and  $W$  the *codomain*<sup>†</sup> for  $L$ .

In particular, setting  $c = 0$  in the second condition implies that a linear function always maps the zero element  $\mathbf{0} \in V$  to the zero element<sup>‡</sup>  $\mathbf{0} \in W$ , so

$$L[\mathbf{0}] = \mathbf{0}. \quad (7.2)$$

We can readily combine the two defining conditions (7.1) into a single rule

$$L[c\mathbf{v} + d\mathbf{w}] = cL[\mathbf{v}] + dL[\mathbf{w}], \quad \text{for all } \mathbf{v}, \mathbf{w} \in V, \quad c, d \in \mathbb{R}, \quad (7.3)$$

that characterizes linearity of a function  $L$ . An easy induction proves that a linear function respects linear combinations, so

$$L[c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k] = c_1L[\mathbf{v}_1] + \cdots + c_kL[\mathbf{v}_k] \quad (7.4)$$

for all  $c_1, \dots, c_k \in \mathbb{R}$  and  $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ .

The interchangeable terms *linear map*, *linear operator*, and, when  $V = W$ , *linear transformation* are all commonly used as alternatives to “linear function”, depending on the

<sup>†</sup> The terms “range” and “target” are also sometimes used for the codomain. However, some authors use “range” to mean the image of  $L$ . An alternative name for domain is “source”.

<sup>‡</sup> We will use the same notation for these two zero elements even though they may belong to different vector spaces. The reader should be able to determine where each lives from the context.

circumstances and taste of the author. The term “linear operator” is particularly useful when the underlying vector space is a function space, so as to avoid confusing the two different uses of the word “function”. As usual, we will often refer to the elements of a vector space as “vectors”, even though they might be functions or matrices or something else, depending on the context.

**Example 7.2.** The simplest linear function is the zero function  $O[\mathbf{v}] \equiv \mathbf{0}$ , which maps every element  $\mathbf{v} \in V$  to the zero vector in  $W$ . Note that, in view of (7.2), this is the *only* constant linear function; a nonzero constant function is *not*, despite its evident simplicity, linear. Another simple but important linear function is the identity function  $I = I_V: V \rightarrow V$ , which maps  $V$  to itself and leaves every vector unchanged:  $I[\mathbf{v}] = \mathbf{v}$ . Slightly more generally, the operation of scalar multiplication  $M_a[\mathbf{v}] = a\mathbf{v}$  by a scalar  $a \in \mathbb{R}$  defines a linear function from  $V$  to itself, with  $M_0 = O$ , the zero function from  $V$  to itself, and  $M_1 = I$ , the identity function on  $V$ , appearing as special cases.

**Example 7.3.** Suppose  $V = \mathbb{R}$ . We claim that every linear function  $L: \mathbb{R} \rightarrow \mathbb{R}$  has the form

$$y = L[x] = ax,$$

for some constant  $a$ . Therefore, the only scalar linear functions are those whose graph is a straight line passing through the origin. To prove this, we write  $x \in \mathbb{R}$  as a scalar product  $x = x \cdot 1$ . Then, by the second property in (7.1),

$$L[x] = L[x \cdot 1] = x \cdot L[1] = ax, \quad \text{where } a = L[1],$$

as claimed.

**Warning.** Even though the graph of the function

$$y = ax + b, \tag{7.5}$$

is a straight line, it is *not* a linear function — unless  $b = 0$ , so the line goes through the origin. The proper mathematical name for a function of the form (7.5) is an *affine function*; see Definition 7.21 below.

**Example 7.4.** Let  $V = \mathbb{R}^n$  and  $W = \mathbb{R}^m$ . Let  $A$  be an  $m \times n$  matrix. Then the function  $L[\mathbf{v}] = A\mathbf{v}$  given by matrix multiplication is easily seen to be a linear function. Indeed, the requirements (7.1) reduce to the basic distributivity and scalar multiplication properties of matrix multiplication:

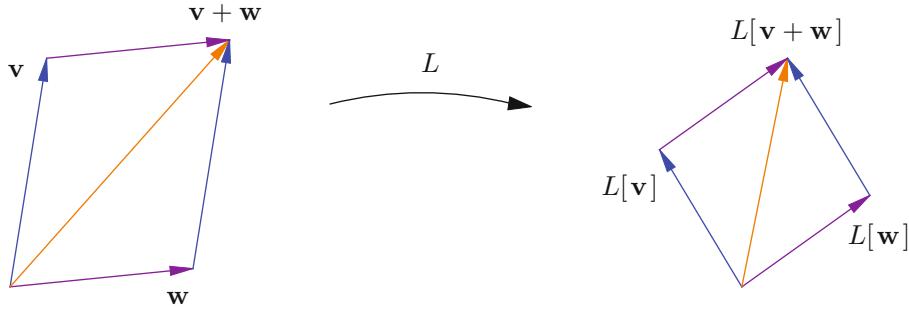
$$A(\mathbf{v} + \mathbf{w}) = A\mathbf{v} + A\mathbf{w}, \quad A(c\mathbf{v}) = cA\mathbf{v}, \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n, \quad c \in \mathbb{R}.$$

In fact, *every* linear function between two Euclidean spaces has this form.

**Theorem 7.5.** Every linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is given by matrix multiplication,  $L[\mathbf{v}] = A\mathbf{v}$ , where  $A$  is an  $m \times n$  matrix.

**Warning.** Pay attention to the order of  $m$  and  $n$ . While  $A$  has size  $m \times n$ , the linear function  $L$  goes *from*  $\mathbb{R}^n$  *to*  $\mathbb{R}^m$ .

*Proof:* The key idea is to look at what the linear function does to the basis vectors. Let  $\mathbf{e}_1, \dots, \mathbf{e}_n$  be the standard basis of  $\mathbb{R}^n$ , as in (2.17), and let  $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_m$  be the standard



**Figure 7.1.** Linear Function on Euclidean Space.

basis of  $\mathbb{R}^m$ . (We temporarily place hats on the latter to avoid confusing the two.) Since  $L[\mathbf{e}_j] \in \mathbb{R}^m$ , we can write it as a linear combination of the latter basis vectors:

$$L[\mathbf{e}_j] = \mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} = a_{1j} \hat{\mathbf{e}}_1 + a_{2j} \hat{\mathbf{e}}_2 + \cdots + a_{mj} \hat{\mathbf{e}}_m, \quad j = 1, \dots, n. \quad (7.6)$$

Let us construct the  $m \times n$  matrix

$$A = (\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_n) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \quad (7.7)$$

whose columns are the image vectors (7.6). Using (7.4), we then compute the effect of  $L$  on a general vector  $\mathbf{v} = (v_1, v_2, \dots, v_n)^T \in \mathbb{R}^n$ :

$$L[\mathbf{v}] = L[v_1 \mathbf{e}_1 + \cdots + v_n \mathbf{e}_n] = v_1 L[\mathbf{e}_1] + \cdots + v_n L[\mathbf{e}_n] = v_1 \mathbf{a}_1 + \cdots + v_n \mathbf{a}_n = A\mathbf{v}.$$

The final equality follows from our basic formula (2.13) connecting matrix multiplication and linear combinations. We conclude that the vector  $L[\mathbf{v}]$  coincides with the vector  $A\mathbf{v}$  obtained by multiplying  $\mathbf{v}$  by the coefficient matrix  $A$ . *Q.E.D.*

The proof of Theorem 7.5 shows us how to construct the matrix representative of a given linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . We merely assemble the image column vectors  $\mathbf{a}_1 = L[\mathbf{e}_1], \dots, \mathbf{a}_n = L[\mathbf{e}_n]$  into an  $m \times n$  matrix  $A$ .

The two basic linearity conditions (7.1) have a simple geometrical interpretation. Since vector addition is the same as completing the parallelogram sketched in Figure 7.1, the first linearity condition requires that  $L$  map parallelograms to parallelograms. The second linearity condition says that if we stretch a vector by a factor  $c$ , then its image under  $L$  must also be stretched by the same amount. Thus, one can often detect linearity by simply looking at the geometry of the function.

**Example 7.6.** As a specific example, consider the function  $R_\theta: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  that rotates the vectors in the plane around the origin by a specified angle  $\theta$ . This geometric transformation clearly preserves parallelograms — see Figure 7.2. It also respects stretching of vectors, and hence defines a linear function. In order to find its matrix representative,

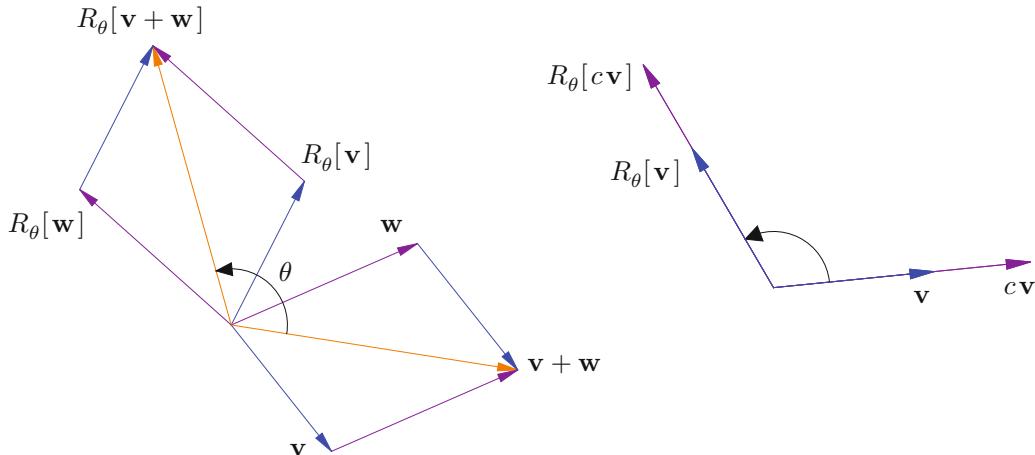
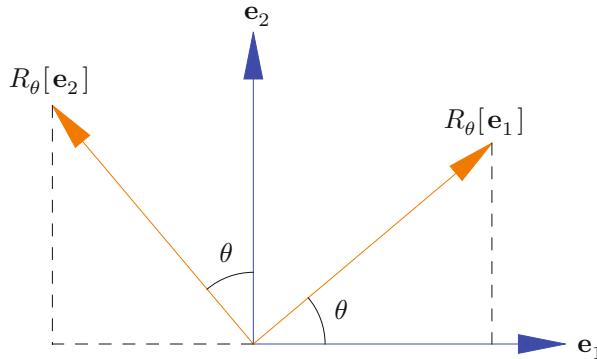


Figure 7.2. Linearity of Rotations.

Figure 7.3. Rotation in  $\mathbb{R}^2$ .

we need to find out where the standard basis vectors  $\mathbf{e}_1, \mathbf{e}_2$  are mapped. Referring to Figure 7.3, and keeping in mind that the rotated vectors also have unit length, we have

$$R_\theta[\mathbf{e}_1] = (\cos \theta) \mathbf{e}_1 + (\sin \theta) \mathbf{e}_2 = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix},$$

$$R_\theta[\mathbf{e}_2] = -(\sin \theta) \mathbf{e}_1 + (\cos \theta) \mathbf{e}_2 = \begin{pmatrix} -\sin \theta \\ \cos \theta \end{pmatrix}.$$

According to the general recipe (7.7), we assemble these two column vectors to obtain the matrix form of the rotation transformation, and so

$$R_\theta[\mathbf{v}] = A_\theta \mathbf{v}, \quad \text{where} \quad A_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}. \quad (7.8)$$

Therefore, rotating a vector  $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$  through angle  $\theta$  produces the vector

$$\hat{\mathbf{v}} = R_\theta[\mathbf{v}] = A_\theta \mathbf{v} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \theta - y \sin \theta \\ x \sin \theta + y \cos \theta \end{pmatrix}$$

with coordinates  $\hat{x} = x \cos \theta - y \sin \theta$ ,  $\hat{y} = x \sin \theta + y \cos \theta$ . These formulas can be proved directly, but, in fact, are a consequence of the underlying linearity of rotations.

## Exercises

7.1.1. Which of the following functions  $F: \mathbb{R}^3 \rightarrow \mathbb{R}$  are linear? (a)  $F(x, y, z) = x$ , (b)  $F(x, y, z) = y - 2$ , (c)  $F(x, y, z) = x + y + 3$ , (d)  $F(x, y, z) = x - y - z$ , (e)  $F(x, y, z) = xyz$ , (f)  $F(x, y, z) = x^2 - y^2 + z^2$ , (g)  $F(x, y, z) = e^{x-y+z}$ .

7.1.2. Explain why the following functions  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  are not linear.

$$(a) \begin{pmatrix} x+2 \\ x+y \end{pmatrix}, \quad (b) \begin{pmatrix} x^2 \\ y^2 \end{pmatrix}, \quad (c) \begin{pmatrix} |y| \\ |x| \end{pmatrix}, \quad (d) \begin{pmatrix} \sin(x+y) \\ x-y \end{pmatrix}, \quad (e) \begin{pmatrix} x+e^y \\ 2x+y \end{pmatrix}.$$

7.1.3. Which of the following functions  $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  are linear?

$$(a) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x-y \\ x+y \end{pmatrix}, \quad (b) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y+1 \\ x-y-1 \end{pmatrix}, \quad (c) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xy \\ x-y \end{pmatrix},$$

$$(d) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3y \\ 2x \end{pmatrix}, \quad (e) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^2+y^2 \\ x^2-y^2 \end{pmatrix}, \quad (f) F\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y-3x \\ x \end{pmatrix}.$$

7.1.4. Explain why the translation function  $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , defined by  $T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+a \\ y+b \end{pmatrix}$  for  $a, b \in \mathbb{R}$ , is almost never linear. Precisely when is it linear?

7.1.5. Find a matrix representation for the following linear transformations on  $\mathbb{R}^3$ :

- (a) counterclockwise rotation by  $90^\circ$  around the  $z$ -axis; (b) clockwise rotation by  $60^\circ$  around the  $x$ -axis; (c) reflection through the  $(x, y)$ -plane; (d) counterclockwise rotation by  $120^\circ$  around the line  $x = y = z$ ; (e) rotation by  $180^\circ$  around the line  $x = y = z$ ; (f) orthogonal projection onto the  $xy$ -plane; (g) orthogonal projection onto the plane  $x - y + 2z = 0$ .

7.1.6. Find a linear function  $L: \mathbb{R}^2 \rightarrow \mathbb{R}$  such that  $L\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2$  and  $L\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 3$ . Is it unique?

7.1.7. Find a linear function  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that  $L\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$  and  $L\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ .

7.1.8. Under what conditions does there exist a linear function  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that

$L\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} a_1 \\ b_1 \end{pmatrix}$  and  $L\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_2 \\ b_2 \end{pmatrix}$ ? Under what conditions is  $L$  uniquely defined? In the latter case, write down the matrix representation of  $L$ .

7.1.9. Can you construct a linear function  $L: \mathbb{R}^3 \rightarrow \mathbb{R}$  such that

$$L\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = 1, \quad L\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} = 4, \quad \text{and} \quad L\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = -2? \quad \text{If yes, find one. If not, explain why not.}$$

◇ 7.1.10. Given  $\mathbf{a} = (a, b, c)^T \in \mathbb{R}^3$ , prove that the cross product map  $L_{\mathbf{a}}[\mathbf{v}] = \mathbf{a} \times \mathbf{v}$ , as defined in (4.2), is linear, and find its matrix representative.

7.1.11. Is the Euclidean norm function  $N(\mathbf{v}) = \|\mathbf{v}\|$ , for  $\mathbf{v} \in \mathbb{R}^n$ , linear?

7.1.12. Let  $V$  be a vector space. Prove that every linear function  $L: \mathbb{R} \rightarrow V$  has the form  $L[x] = x \mathbf{b}$ , where  $x \in \mathbb{R}$ , for some  $\mathbf{b} \in V$ .

7.1.13. *True or false:* The quadratic form  $Q(\mathbf{v}) = \mathbf{v}^T K \mathbf{v}$  defined by a symmetric  $n \times n$  matrix  $K$  defines a linear function  $Q: \mathbb{R}^n \rightarrow \mathbb{R}$ .

◇ 7.1.14. (a) Prove that  $L$  is linear if and only if it satisfies (7.3).

(b) Use induction to prove that  $L$  satisfies (7.4).

- 7.1.15. Let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ ,  $B = \begin{pmatrix} p & q \\ r & s \end{pmatrix}$  be  $2 \times 2$  matrices. For each of the following functions, prove that  $L: \mathcal{M}_{2 \times 2} \rightarrow \mathcal{M}_{2 \times 2}$  defines a linear map, and then find its matrix representative with respect to the standard basis  $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$  of  $\mathcal{M}_{2 \times 2}$ :
- (a)  $L[X] = AX$ , (b)  $R[X] = XB$ , (c)  $K[X] = AXB$ .

- 7.1.16. The domain space of the following functions is the space of  $n \times n$  real matrices  $A$ .

- Which are linear? What is the codomain space in each case?
- (a)  $L[A] = 3A$ ;  
 (b)  $L[A] = I - A$ ;  
 (c)  $L[A] = A^T$ ;  
 (d)  $L[A] = A^{-1}$ ;  
 (e)  $L[A] = \det A$ ;  
 (f)  $L[A] = \text{tr } A$ ;  
 (g)  $L[A] = (a_{11}, \dots, a_{nn})^T$ , i.e., the vector of diagonal entries of  $A$ ;  
 (h)  $L[A] = A\mathbf{v}$ , where  $\mathbf{v} \in \mathbb{R}^n$ ;  
 (i)  $L[A] = \mathbf{v}^T A \mathbf{v}$ , where  $\mathbf{v} \in \mathbb{R}^n$ .

- ◇ 7.1.17. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis of  $V$  and  $\mathbf{w}_1, \dots, \mathbf{w}_n$  be any vectors in  $W$ . Show that there is a unique linear function  $L: V \rightarrow W$  such that  $L[\mathbf{v}_i] = \mathbf{w}_i$ ,  $i = 1, \dots, n$ .

- ◇ 7.1.18. *Bilinear functions:* Let  $V, W, Z$  be vector spaces. A function that takes any pair of vectors  $\mathbf{v} \in V$  and  $\mathbf{w} \in W$  to a vector  $\mathbf{z} = B[\mathbf{v}, \mathbf{w}] \in Z$  is called *bilinear* if, for each fixed  $\mathbf{w}$ , it is a linear function of  $\mathbf{v}$ , so  $B[c\mathbf{v} + d\tilde{\mathbf{v}}, \mathbf{w}] = cB[\mathbf{v}, \mathbf{w}] + dB[\tilde{\mathbf{v}}, \mathbf{w}]$ , and, for each fixed  $\mathbf{v}$ , it is a linear function of  $\mathbf{w}$ , so  $B[\mathbf{v}, c\mathbf{w} + d\tilde{\mathbf{w}}] = cB[\mathbf{v}, \mathbf{w}] + dB[\mathbf{v}, \tilde{\mathbf{w}}]$ . Thus,  $B: V \times W \rightarrow Z$  defines a function on the Cartesian product space  $V \times W$ , as defined in Exercise 2.1.13. (a) Show that  $B[\mathbf{v}, \mathbf{w}] = v_1 w_1 - 2v_2 w_2$  is a bilinear function from  $\mathbb{R}^2 \times \mathbb{R}^2$  to  $\mathbb{R}$ . (b) Show that  $B[\mathbf{v}, \mathbf{w}] = 2v_1 w_2 - 3v_2 w_3$  is a bilinear function from  $\mathbb{R}^2 \times \mathbb{R}^3$  to  $\mathbb{R}$ . (c) Show that if  $V$  is an inner product space, then  $B[\mathbf{v}, \mathbf{w}] = \langle \mathbf{v}, \mathbf{w} \rangle$  defines a bilinear function  $B: V \times V \rightarrow \mathbb{R}$ . (d) Show that if  $A$  is any  $m \times n$  matrix, then  $B[\mathbf{v}, \mathbf{w}] = \mathbf{v}^T A \mathbf{w}$  defines a bilinear function  $B: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ . (e) Show that every bilinear function  $B: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$  arises in this way. (f) Show that a vector-valued function  $B: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^k$  defines a bilinear function if and only if each of its components  $B_i: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$ , for  $i = 1, \dots, k$ , is a bilinear function. (g) *True or false:* A bilinear function  $B: V \times W \rightarrow Z$  defines a linear function on the Cartesian product space.

## Linear Operators

So far, we have concentrated on linear functions on Euclidean space, and discovered that they are all represented by matrices. For function spaces, there is a much wider variety of linear operators available, and a complete classification is out of the question. Let us look at some of the main representative examples that arise in applications.

**Example 7.7.** (a) Recall that  $C^0[a, b]$  denotes the vector space consisting of all continuous functions on the interval  $[a, b]$ . Evaluation of the function at a point, namely  $L[f] = f(x_0)$ , defines a linear operator  $L: C^0[a, b] \rightarrow \mathbb{R}$ , because

$$L[cf + dg] = (cf + dg)(x_0) = c f(x_0) + d g(x_0) = c L[f] + d L[g]$$

for any functions  $f, g \in C^0[a, b]$  and scalars (constants)  $c, d$ .

(b) Another real-valued linear function is the integration operator

$$I[f] = \int_a^b f(x) dx, \tag{7.9}$$

that maps  $I: C^0[a, b] \rightarrow \mathbb{R}$ . Linearity of  $I$  is an immediate consequence of the basic integration identity

$$\int_a^b [cf(x) + dg(x)] dx = c \int_a^b f(x) dx + d \int_a^b g(x) dx,$$

which is valid for arbitrary integrable — which includes continuous — functions  $f, g$  and constants  $c, d$ .

(c) We have already seen that multiplication of functions by a constant,  $M_c[f(x)] = cf(x)$ , defines a linear map  $M_c: C^0[a, b] \rightarrow C^0[a, b]$ ; the particular case  $c = 1$  reduces to the identity transformation  $I = M_1$ . More generally, if  $a(x) \in C^0[a, b]$  is a given continuous function, then the operation  $M_a[f(x)] = a(x)f(x)$  of multiplication by  $a$  also defines a linear transformation  $M_a: C^0[a, b] \rightarrow C^0[a, b]$ .

(d) Another important linear transformation is the indefinite integral

$$J[f] = g, \quad \text{where} \quad g(x) = \int_a^x f(y) dy. \quad (7.10)$$

According to the Fundamental Theorem of Calculus, [2, 78], the integral of a continuous function is continuously differentiable. Therefore,  $J: C^0[a, b] \rightarrow C^1[a, b]$  defines a linear operator from the space of continuous functions to the space of continuously differentiable functions.

(e) Conversely, differentiation of functions is also a linear operation. To be precise, since not every continuous function can be differentiated, we take the domain space to be the vector space  $C^1[a, b]$  of continuously differentiable functions on the interval  $[a, b]$ . The derivative operator

$$D[f] = f' \quad (7.11)$$

defines a linear operator  $D: C^1[a, b] \rightarrow C^0[a, b]$ . This follows from the elementary differentiation formula

$$D[cf + dg] = (cf + dg)' = cf' + dg' = cD[f] + dD[g],$$

valid whenever  $c, d$  are constant.

## Exercises

7.1.19. Which of the following define linear operators on the vector space  $C^1(\mathbb{R})$  of continuously differentiable scalar functions? What is the codomain?

- (a)  $L[f] = f(0) + f(1)$ , (b)  $L[f] = f(0)f(1)$ , (c)  $L[f] = f'(1)$ , (d)  $L[f] = f'(3) - f(2)$ ,
- (e)  $L[f] = x^2 f(x)$ , (f)  $L[f] = f(x+2)$ , (g)  $L[f] = f(x) + 2$ , (h)  $L[f] = f'(2x)$ ,
- (i)  $L[f] = f'(x^2)$ , (j)  $L[f] = f(x)\sin x - f'(x)\cos x$ , (k)  $L[f] = 2 \log f(0)$ ,
- (l)  $L[f] = \int_0^1 e^y f(y) dy$ , (m)  $L[f] = \int_0^1 |f(y)| dy$ , (n)  $L[f] = \int_{x-1}^{x+1} f(y) dy$ ,
- (o)  $L[f] = \int_x^{x^2} \frac{f(y)}{y} dy$ , (p)  $L[f] = \int_0^{f(x)} y dy$ , (q)  $L[f] = \int_0^x y^2 f'(y) dy$ ,
- (r)  $L[f] = \int_{-1}^1 [f(y) - f(0)] dy$ , (s)  $L[f] = \int_{-1}^x [f(y) - y] dy$ .

7.1.20. *True or false:* The average or mean  $A[f] = \frac{1}{b-a} \int_a^b f(x) dx$  of a function on the interval  $[a, b]$  defines a linear operator  $A: C^0[a, b] \rightarrow \mathbb{R}$ .

7.1.21. Prove that multiplication  $M_h[f(x)] = h(x)f(x)$  by a given function  $h \in C^n[a, b]$  defines a linear operator  $M_h: C^n[a, b] \rightarrow C^n[a, b]$ . Which result from calculus do you need to complete the proof?

7.1.22. Show that if  $w(x)$  is any continuous function, then the weighted integral

$$I_w[f] = \int_a^b f(x) w(x) dx$$

defines a linear operator  $I_w: C^0[a, b] \rightarrow \mathbb{R}$ .

7.1.23. (a) Show that the partial derivatives  $\partial_x[f] = \frac{\partial f}{\partial x}$  and  $\partial_y[f] = \frac{\partial f}{\partial y}$  both define linear operators on the space of continuously differentiable functions  $f(x, y)$ .

(b) For which values of  $a, b, c, d$  is the map  $L[f] = a \frac{\partial f}{\partial x} + b \frac{\partial f}{\partial y} + c f + d$  linear?

7.1.24. Prove that the Laplacian operator  $\Delta[f] = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$  defines a linear function on the vector space of twice continuously differentiable functions  $f(x, y)$ .

7.1.25. Show that the gradient  $G[f] = \nabla f$  defines a linear operator from the space of continuously differentiable scalar-valued functions  $f: \mathbb{R}^2 \rightarrow \mathbb{R}$  to the space of continuous vector fields  $\mathbf{v}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ .

7.1.26. Prove that, on  $\mathbb{R}^3$ , the gradient, curl, and divergence all define linear operators. Be precise in your description of the domain space and the codomain space in each case.

## The Space of Linear Functions

Given two vector spaces  $V, W$ , we use  $\mathcal{L}(V, W)$  to denote the set of all<sup>†</sup> linear functions  $L: V \rightarrow W$ . We claim that  $\mathcal{L}(V, W)$  is itself a vector space. We add linear functions  $L, M \in \mathcal{L}(V, W)$  in the same way we add general functions:

$$(L + M)[\mathbf{v}] = L[\mathbf{v}] + M[\mathbf{v}].$$

You should check that  $L + M$  satisfies the linear function axioms (7.1), provided that  $L$  and  $M$  do. Similarly, multiplication of a linear function by a scalar  $c \in \mathbb{R}$  is defined so that  $(cL)[\mathbf{v}] = cL[\mathbf{v}]$ , again producing a linear function. The zero element of  $\mathcal{L}(V, W)$  is the zero function  $O[\mathbf{v}] \equiv \mathbf{0}$ . The verification that  $\mathcal{L}(V, W)$  satisfies the basic vector space axioms of Definition 2.1 is left to the reader.

In particular, if  $V = \mathbb{R}^n$  and  $W = \mathbb{R}^m$ , then Theorem 7.5 implies that we can identify  $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$  with the space  $\mathcal{M}_{m \times n}$  of all  $m \times n$  matrices. Addition of linear functions corresponds to matrix addition, while scalar multiplication coincides with the usual scalar multiplication of matrices. (Why?) Therefore, the space of all  $m \times n$  matrices is a vector space — a fact we already knew. The *standard basis* for  $\mathcal{M}_{m \times n}$  is given by the  $m n$  matrices  $E_{ij}$ ,  $1 \leq i \leq m$ ,  $1 \leq j \leq n$ , which have a single 1 in the  $(i, j)$  position and zeros everywhere else. Therefore, the dimension of  $\mathcal{M}_{m \times n}$  is  $m n$ . Note that  $E_{ij}$  corresponds to the specific linear transformation that maps  $E_{ij}[\mathbf{e}_j] = \hat{\mathbf{e}}_i$ , while  $E_{ij}[\mathbf{e}_k] = \mathbf{0}$  whenever  $k \neq j$ .

**Example 7.8.** The space of linear transformations of the plane,  $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$ , is identified with the space  $\mathcal{M}_{2 \times 2}$  of  $2 \times 2$  matrices  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . The standard basis of  $\mathcal{M}_{2 \times 2}$  consists

<sup>†</sup> In infinite-dimensional situations, one usually imposes additional restrictions, e.g., continuity or boundedness of the linear operators. We shall relegate these more subtle distinctions to a more advanced treatment of the subject. See [50, 67] for a full discussion of the rather sophisticated analytical details, which play an important role in serious quantum mechanical applications.

of the  $4 = 2 \cdot 2$  matrices

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Indeed, we can uniquely write any other matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = aE_{11} + bE_{12} + cE_{21} + dE_{22},$$

as a linear combination of these four basis matrices. Of course, as with any vector space, this is but one of many other possible bases of  $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$ .

## Dual Spaces

A particularly important case is that in which the codomain of the linear functions is  $\mathbb{R}$ .

**Definition 7.9.** The *dual space* to a vector space  $V$  is the vector space  $V^* = \mathcal{L}(V, \mathbb{R})$  consisting of all real-valued linear functions  $\ell: V \rightarrow \mathbb{R}$ .

If  $V = \mathbb{R}^n$ , then, by Theorem 7.5, every linear function  $\ell: \mathbb{R}^n \rightarrow \mathbb{R}$  is given by multiplication by a  $1 \times n$  matrix, i.e., a row vector. Explicitly,

$$\ell[\mathbf{v}] = \mathbf{a} \mathbf{v} = a_1 v_1 + \cdots + a_n v_n, \quad \text{where} \quad \mathbf{a} = (a_1 \ a_2 \ \dots \ a_n), \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}.$$

Therefore, we can identify the dual space  $(\mathbb{R}^n)^*$  with the space of *row* vectors with  $n$  entries. In light of this observation, the distinction between row vectors and column vectors is now seen to be much more sophisticated than mere semantics or notation. Row vectors should more properly be viewed as real-valued linear functions — the *dual* objects to column vectors.

The *standard dual basis*  $\varepsilon_1, \dots, \varepsilon_n$  of  $(\mathbb{R}^n)^*$  consists of the standard row basis vectors; namely,  $\varepsilon_j$  is the row vector with 1 in the  $j^{\text{th}}$  slot and zeros elsewhere. The  $j^{\text{th}}$  dual basis element defines the linear function

$$E_j[\mathbf{v}] = \varepsilon_j \mathbf{v} = v_j,$$

which picks off the  $j^{\text{th}}$  coordinate of  $\mathbf{v}$  — with respect to the original basis  $\mathbf{e}_1, \dots, \mathbf{e}_n$ . Thus, the dimensions of  $V = \mathbb{R}^n$  and its dual  $V^* = (\mathbb{R}^n)^*$  are both equal to  $n$ .

An inner product structure provides a mechanism for identifying a vector space and its dual. However, it should be borne in mind that this identification will depend upon the choice of inner product.

**Theorem 7.10.** Let  $V$  be a finite-dimensional real inner product space. Then every linear function  $\ell: V \rightarrow \mathbb{R}$  is given by taking the inner product with a fixed vector  $\mathbf{a} \in V$ :

$$\ell[\mathbf{v}] = \langle \mathbf{a}, \mathbf{v} \rangle. \tag{7.12}$$

*Proof:* Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis of  $V$ . If we write  $\mathbf{v} = y_1 \mathbf{v}_1 + \cdots + y_n \mathbf{v}_n$ , then, by linearity,

$$\ell[\mathbf{v}] = y_1 \ell[\mathbf{v}_1] + \cdots + y_n \ell[\mathbf{v}_n] = b_1 y_1 + \cdots + b_n y_n, \quad \text{where} \quad b_i = \ell[\mathbf{v}_i]. \tag{7.13}$$

On the other hand, if we write  $\mathbf{a} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n$ , then

$$\langle \mathbf{a}, \mathbf{v} \rangle = \sum_{i,j=1}^n x_j y_i \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \sum_{i,j=1}^n g_{ij} x_j y_i, \quad (7.14)$$

where  $G = (g_{ij})$  is the  $n \times n$  Gram matrix with entries  $g_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$ . Equality of (7.13, 14) requires that  $G\mathbf{x} = \mathbf{b}$ , where  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ ,  $\mathbf{b} = (b_1, b_2, \dots, b_n)^T$ . Invertibility of  $G$  as guaranteed by Theorem 3.34, allows us to solve for  $\mathbf{x} = G^{-1}\mathbf{b}$  and thereby construct the desired vector  $\mathbf{a}$ . In particular, if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  is an orthonormal basis, then  $G = I$  and hence  $\mathbf{a} = b_1 \mathbf{v}_1 + \cdots + b_n \mathbf{v}_n$ . *Q.E.D.*

**Remark.** For the particular case in which  $V = \mathbb{R}^n$  is endowed with the standard dot product, Theorem 7.10 identifies a row vector representing a linear function with the corresponding column vector obtained by transposition  $\mathbf{a} \mapsto \mathbf{a}^T$ . Thus, the naïve identification of a row and a column vector is, in fact, an indication of a much more subtle phenomenon that relies on the identification of  $\mathbb{R}^n$  with its dual based on the Euclidean inner product. Alternative inner products will lead to alternative, more complicated, identifications of row and column vectors; see Exercise 7.1.31 for details.

**Important.** Theorem 7.10 is *not* true if  $V$  is infinite-dimensional. This fact will have important repercussions for the analysis of the differential equations of continuum mechanics, which will lead us immediately into the much deeper waters of generalized function theory, as described in [61].

## Exercises

7.1.27. Write down a basis for and dimension of the linear function spaces (a)  $\mathcal{L}(\mathbb{R}^3, \mathbb{R})$ , (b)  $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$ , (c)  $\mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ , (d)  $\mathcal{L}(\mathcal{P}^{(3)}, \mathbb{R})$ , (e)  $\mathcal{L}(\mathcal{P}^{(2)}, \mathbb{R}^2)$ , (f)  $\mathcal{L}(\mathcal{P}^{(2)}, \mathcal{P}^{(2)})$ . Here  $\mathcal{P}^{(n)}$  is the space of polynomials of degree  $\leq n$ .

7.1.28. *True or false:* The set of linear transformations  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  such that  $L \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$  is a subspace of  $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$ . If true, what is its dimension?

7.1.29. *True or false:* The set of linear transformations  $L: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  such that  $L \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$  is a subspace of  $\mathcal{L}(\mathbb{R}^3, \mathbb{R}^3)$ . If true, what is its dimension?

7.1.30. Consider the linear function  $L: \mathbb{R}^3 \rightarrow \mathbb{R}$  defined by  $L(x, y, z) = 3x - y + 2z$ . Write down the vector  $\mathbf{a} \in \mathbb{R}^3$  such that  $L[\mathbf{v}] = \langle \mathbf{a}, \mathbf{v} \rangle$  when the inner product is (a) the Euclidean dot product; (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$ ; (c) the inner product defined by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ .

◇ 7.1.31. Let  $\mathbb{R}^n$  be equipped with the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T K \mathbf{w}$ . Let  $L[\mathbf{v}] = \mathbf{r} \mathbf{v}$  where  $\mathbf{r}$  is a row vector of size  $1 \times n$ . (a) Find a formula for the column vector  $\mathbf{a}$  such that (7.12) holds for the linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}$ . (b) Illustrate your result when  $\mathbf{r} = (2, -1)$ , using (i) the dot product (ii) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 3v_1 w_1 + 2v_2 w_2$ , (iii) the inner product induced by  $K = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ .

♡ 7.1.32. *Dual Bases:* Given a basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $V$ , the *dual basis*  $\ell_1, \dots, \ell_n$  of  $V^*$  consists of the linear functions uniquely defined by the requirements  $\ell_i(\mathbf{v}_j) = \begin{cases} 1 & i = j, \\ 0, & i \neq j. \end{cases}$

(a) Show that  $\ell_i[\mathbf{v}] = x_i$  gives the  $i^{\text{th}}$  coordinate of a vector  $\mathbf{v} = x_1\mathbf{v}_1 + \dots + x_n\mathbf{v}_n$  with respect to the given basis. (b) Prove that the dual basis is indeed a basis for the dual vector space. (c) Prove that if  $V = \mathbb{R}^n$  and  $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  is the  $n \times n$  matrix whose columns are the basis vectors, then the rows of the inverse matrix  $A^{-1}$  can be identified as the corresponding dual basis of  $(\mathbb{R}^n)^*$ .

7.1.33. Use Exercise 7.1.32(c) to find the dual basis for: (a)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ; (b)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ ; (c)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ ; (d)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 0 \\ -3 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}$ ; (e)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_4 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$ .

7.1.34. Let  $\mathcal{P}^{(2)}$  denote the space of quadratic polynomials equipped with the  $L^2$  inner product  $\langle p, q \rangle = \int_0^1 p(x)q(x)dx$ . Find the polynomial  $q$  that represents the following linear functions, i.e., such that  $L[p] = \langle q, p \rangle$ : (a)  $L[p] = p(0)$ , (b)  $L[p] = \frac{1}{2}p'(1)$ , (c)  $L[p] = \int_0^1 p(x)dx$ , (d)  $L[p] = \int_{-1}^1 p(x)dx$ .

7.1.35. Find the dual basis, as defined in Exercise 7.1.32, for the monomial basis of  $\mathcal{P}^{(2)}$  with respect to the  $L^2$  inner product  $\langle p, q \rangle = \int_0^1 p(x)q(x)dx$ .

7.1.36. Write out a proof of Theorem 7.10 that does not rely on finding an orthonormal basis.

## Composition

Besides adding and multiplying by scalars, one can also compose linear functions.

**Lemma 7.11.** Let  $V, W, Z$  be vector spaces. If  $L: V \rightarrow W$  and  $M: W \rightarrow Z$  are linear functions, then the composite function  $M \circ L: V \rightarrow Z$ , defined by  $(M \circ L)[\mathbf{v}] = M[L[\mathbf{v}]]$  is also linear.

*Proof:* This is straightforward:

$$\begin{aligned} (M \circ L)[c\mathbf{v} + d\mathbf{w}] &= M[L[c\mathbf{v} + d\mathbf{w}]] = M[cL[\mathbf{v}] + dL[\mathbf{w}]] \\ &= cM[L[\mathbf{v}]] + dM[L[\mathbf{w}]] = c(M \circ L)[\mathbf{v}] + d(M \circ L)[\mathbf{w}], \end{aligned}$$

where we used, successively, the linearity of  $L$  and then of  $M$ .

Q.E.D.

For example, if  $L[\mathbf{v}] = A\mathbf{v}$  maps  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and  $M[\mathbf{w}] = B\mathbf{w}$  maps  $\mathbb{R}^m$  to  $\mathbb{R}^l$ , so that  $A$  is an  $m \times n$  matrix and  $B$  is a  $l \times m$  matrix, then

$$(M \circ L)[\mathbf{v}] = M[L[\mathbf{v}]] = B(A\mathbf{v}) = (BA)\mathbf{v},$$

and hence the composition  $M \circ L: \mathbb{R}^n \rightarrow \mathbb{R}^l$  corresponds to the  $l \times n$  product matrix  $BA$ . In other words, on Euclidean space, *composition of linear functions is the same as matrix multiplication*. And, like matrix multiplication, composition of (linear) functions is not, in general, commutative.

**Example 7.12.** Composing two rotations results in another rotation:  $R_\varphi \circ R_\theta = R_{\varphi+\theta}$ . In other words, if we first rotate by angle  $\theta$  and then by angle  $\varphi$ , the net effect is rotation by angle  $\varphi + \theta$ . On the matrix level of (7.8), this implies that

$$\begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} = A_\varphi A_\theta = A_{\varphi+\theta} = \begin{pmatrix} \cos(\varphi + \theta) & -\sin(\varphi + \theta) \\ \sin(\varphi + \theta) & \cos(\varphi + \theta) \end{pmatrix}.$$

Multiplying out the left-hand side, we deduce the well-known trigonometric addition formulas

$$\cos(\varphi + \theta) = \cos \varphi \cos \theta - \sin \varphi \sin \theta, \quad \sin(\varphi + \theta) = \cos \varphi \sin \theta + \sin \varphi \cos \theta.$$

In fact, this constitutes a *bona fide* proof of these two trigonometric identities!

**Example 7.13.** One can build up more sophisticated linear operators on function space by adding and composing simpler ones. In particular, higher order derivative operators are obtained by composing the derivative operator  $D$ , defined in (7.11), with itself. For example,

$$D^2[f] = D \circ D[f] = D[f'] = f''$$

defines the second derivative operator. One needs to exercise due care about the domain of definition, since not every function is differentiable. In general, the  $k^{\text{th}}$  order derivative

$D^k[f] = f^{(k)}(x)$  defines a linear operator  $D^k: C^n[a, b] \rightarrow C^{n-k}[a, b]$  for all  $n \geq k$ ,

obtained by composing  $D$  with itself  $k$  times.

If we further compose  $D^k$  with the linear operation of multiplication by a given function  $a(x)$  we obtain the linear operator  $(a D^k)[f] = a(x) f^{(k)}(x)$ . Finally, a general *linear ordinary differential operator* of order  $n$ ,

$$L = a_n(x) D^n + a_{n-1}(x) D^{n-1} + \cdots + a_1(x) D + a_0(x), \quad (7.15)$$

is obtained by summing such operators. If the coefficient functions  $a_0(x), \dots, a_n(x)$  are continuous, then

$$L[u] = a_n(x) \frac{d^n u}{dx^n} + a_{n-1}(x) \frac{d^{n-1} u}{dx^{n-1}} + \cdots + a_1(x) \frac{du}{dx} + a_0(x)u \quad (7.16)$$

defines a linear operator from  $C^n[a, b]$  to  $C^0[a, b]$ . The most important case — but certainly not the only one arising in applications — is when the coefficients  $a_i(x) = c_i$  are all constant.

## Exercises

7.1.37. For each of the following pairs of linear functions  $S, T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , describe the compositions  $S \circ T$  and  $T \circ S$ . Do the functions commute?

- (a)  $S$  = counterclockwise rotation by  $60^\circ$ ;  $T$  = clockwise rotation by  $120^\circ$ ;
- (b)  $S$  = reflection in the line  $y = x$ ;  $T$  = rotation by  $180^\circ$ ;
- (c)  $S$  = reflection in the  $x$ -axis;  $T$  = reflection in the  $y$ -axis;
- (d)  $S$  = reflection in the line  $y = x$ ;  $T$  = reflection in the line  $y = 2x$ ;
- (e)  $S$  = orthogonal projection on the  $x$ -axis;  $T$  = orthogonal projection on the  $y$ -axis;
- (f)  $S$  = orthogonal projection on the  $x$ -axis;  $T$  = orthogonal projection on the line  $y = x$ ;
- (g)  $S$  = orthogonal projection on the  $x$ -axis;  $T$  = rotation by  $180^\circ$ ;
- (h)  $S$  = orthogonal projection on the  $x$ -axis;  $T$  = counterclockwise rotation by  $90^\circ$ ;
- (i)  $S$  = orthogonal projection on the line  $y = -2x$ ;  $T$  = reflection in the line  $y = x$ .

- 7.1.38. Find a matrix representative for the linear functions (a)  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  that maps  $\mathbf{e}_1$  to  $\begin{pmatrix} 1 \\ -3 \end{pmatrix}$  and  $\mathbf{e}_2$  to  $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$ ; (b)  $M: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  that takes  $\mathbf{e}_1$  to  $\begin{pmatrix} -1 \\ -3 \end{pmatrix}$  and  $\mathbf{e}_2$  to  $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ ; and (c)  $N: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  that takes  $\begin{pmatrix} 1 \\ -3 \end{pmatrix}$  to  $\begin{pmatrix} -1 \\ -3 \end{pmatrix}$  and  $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$  to  $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ . (d) Explain why  $M = N \circ L$ . (e) Verify part (d) by multiplying the matrix representatives.

- 7.1.39. On the vector space  $\mathbb{R}^3$ , let  $R$  denote counterclockwise rotation around the  $x$  axis by  $90^\circ$  and  $S$  counterclockwise rotation around the  $z$ -axis by  $90^\circ$ . (a) Find matrix representatives for  $R$  and  $S$ . (b) Show that  $R \circ S \neq S \circ R$ . Explain what happens to the standard basis vectors under the two compositions. (c) Give an experimental demonstration of the noncommutativity of  $R$  and  $S$  by physically rotating a solid object, e.g., this book, in the prescribed manners.

- 7.1.40. Let  $P$  denote orthogonal projection of  $\mathbb{R}^3$  onto the plane  $V = \{z = x + y\}$  and  $Q$  denote orthogonal projection onto the plane  $W = \{z = x - y\}$ . Is the composition  $R = Q \circ P$  the same as orthogonal projection onto the line  $L = V \cap W$ ? Verify your conclusion by computing the matrix representatives of  $P, Q$ , and  $R$ .

- 7.1.41. (a) Write the linear operator  $L[f(x)] = f'(b)$  as a composition of two linear functions. Do your linear functions commute? (b) For which values of  $a, b, c, d, e$  is  $L[f(x)] = a f'(b) + c f(d) + e$  a linear function?

- 7.1.42. Let  $L = xD + 1$ , and  $M = D - x$  be differential operators. Find  $L \circ M$  and  $M \circ L$ . Do the differential operators commute?

- 7.1.43. Show that the space of constant coefficient linear differential operators of order  $\leq n$  is a vector space. Determine its dimension by exhibiting a basis.

- 7.1.44. (a) Explain why the differential operator  $L = D \circ M_a \circ D$  obtained by composing the linear operators of differentiation  $D[f(x)] = f'(x)$  and multiplication  $M_a[f(x)] = a(x)f(x)$  by a given function  $a(x)$  defines a linear operator. (b) Re-express  $L$  as a linear differential operator of the form (7.16).

- ◇ 7.1.45. (a) Show that composition of linear functions is associative:  $(L \circ M) \circ N = L \circ (M \circ N)$ . Be precise about the domain and codomain spaces involved. (b) How do you know the result is a linear function? (c) Explain why this proves associativity of matrix multiplication.

- 7.1.46. Show that if  $p(x, y)$  is any polynomial, then  $L = p(\partial_x, \partial_y)$  defines a linear, constant coefficient partial differential operator. For example, if  $p(x, y) = x^2 + y^2$ , then  $L = \partial_x^2 + \partial_y^2$  is the Laplacian operator  $\Delta[f] = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$ .

- ◇ 7.1.47. The *commutator* of two linear transformations  $L, M: V \rightarrow V$  on a vector space  $V$  is

$$K = [L, M] = L \circ M - M \circ L. \quad (7.17)$$

- (a) Prove that the commutator  $K$  is a linear transformation on  $V$ . (b) Explain why Exercise 1.2.30 is a special case. (c) Prove that  $L$  and  $M$  commute if and only if  $[L, M] = 0$ . (d) Compute the commutators of the linear transformations defined by the following pairs of matrices:

$$(i) \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 1 & 2 \end{pmatrix}, \quad (ii) \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (iii) \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

- (e) Prove that the *Jacobi identity*

$$[[L, M], N] + [[N, L], M] + [[M, N], L] = 0 \quad (7.18)$$

- is valid for any three linear transformations. (f) Verify the Jacobi identity for the first three matrices in part (c). (g) Prove that the commutator  $B[L, M] = [L, M]$  defines a bilinear map  $B: \mathcal{L}(V, V) \times \mathcal{L}(V, V) \rightarrow \mathcal{L}(V, V)$  on the Cartesian product space, cf. Exercise 7.1.18.

- ◇ 7.1.48. (a) In (one-dimensional) quantum mechanics, the differentiation operator  $P[f(x)] = f'(x)$  represents the *momentum* of a particle, while the operator  $Q[f(x)] = xf(x)$  of multiplication by the function  $x$  represents its *position*. Prove that the position and momentum operators satisfy the *Heisenberg Commutation Relations*  $[P, Q] = P \circ Q - Q \circ P = I$ . (b) Prove that there are no matrices  $P, Q$  that satisfy the Heisenberg Commutation Relations. *Hint:* Use Exercise 1.2.31.
- Remark.** The noncommutativity of quantum mechanical observables lies at the heart of the Uncertainty Principle. The result in part (b) is one of the main reasons why quantum mechanics must be an intrinsically infinite-dimensional theory.
- ◇ 7.1.49. Let  $\mathcal{D}^{(1)}$  denote the set of all first order linear differential operators  $L = p(x)D + q(x)$  where  $p, q$  are polynomials. (a) Prove that  $\mathcal{D}^{(1)}$  is a vector space. Is it finite-dimensional or infinite-dimensional? (b) Prove that the commutator (7.17) of  $L, M \in \mathcal{D}^{(1)}$  is a first order differential operator  $[L, M] \in \mathcal{D}^{(1)}$  by writing out an explicit formula. (c) Verify the Jacobi identity (7.18) for the first order operators  $L = D$ ,  $M = xD + 1$ , and  $N = x^2D + 2x$ .
- 7.1.50. Do the conclusions of Exercise 7.1.49(a-b) hold for the space  $\mathcal{D}^{(2)}$  of second order differential operators  $L = p(x)D^2 + q(x)D + r(x)$ , where  $p, q, r$  are polynomials?

## Inverses

The inverse of a linear function is defined in direct analogy with the Definition 1.13 of the inverse of a (square) matrix.

**Definition 7.14.** Let  $L: V \rightarrow W$  be a linear function. If  $M: W \rightarrow V$  is a function such that both compositions

$$L \circ M = I_W, \quad M \circ L = I_V, \quad (7.19)$$

are equal to the identity function, then we call  $M$  the *inverse* of  $L$  and write  $M = L^{-1}$ .

The two conditions (7.19) require

$$L[M[\mathbf{w}]] = \mathbf{w} \quad \text{for all } \mathbf{w} \in W, \quad \text{and} \quad M[L[\mathbf{v}]] = \mathbf{v} \quad \text{for all } \mathbf{v} \in V.$$

In Exercise 7.1.55, you are asked to prove that, when it exists, the inverse is unique. Of course, if  $M = L^{-1}$  is the inverse of  $L$ , then  $L = M^{-1}$  is the inverse of  $M$  since the conditions are symmetric, and, in such cases,  $(L^{-1})^{-1} = L$ .

**Lemma 7.15.** If it exists, the inverse of a linear function is also a linear function.

*Proof:* Let  $L, M$  satisfy the conditions of Definition 7.14. Given  $\mathbf{w}, \tilde{\mathbf{w}} \in W$ , we note

$$\mathbf{w} = (L \circ M)[\mathbf{w}] = L[\mathbf{v}], \quad \tilde{\mathbf{w}} = (L \circ M)[\tilde{\mathbf{w}}] = L[\tilde{\mathbf{v}}], \quad \text{where } \mathbf{v} = M[\mathbf{w}], \quad \tilde{\mathbf{v}} = M[\tilde{\mathbf{w}}].$$

Therefore, given scalars  $c, d$ , and using only the linearity of  $L$ ,

$$M[c\mathbf{w} + d\tilde{\mathbf{w}}] = M[cL[\mathbf{v}] + dL[\tilde{\mathbf{v}}]] = (M \circ L)[c\mathbf{v} + d\tilde{\mathbf{v}}] = c\mathbf{v} + d\tilde{\mathbf{v}} = cM[\mathbf{w}] + dM[\tilde{\mathbf{w}}],$$

proving linearity of  $M$ . *Q.E.D.*

If  $V = \mathbb{R}^n$ ,  $W = \mathbb{R}^m$ , so that  $L$  and  $M$  are given by matrix multiplication, by  $A$  and  $B$  respectively, then the conditions (7.19) reduce to the usual conditions

$$AB = I, \quad BA = I,$$

for matrix inversion, cf. (1.37). Therefore,  $B = A^{-1}$  is the inverse matrix. In particular, for  $L: \mathbb{R}^m \rightarrow \mathbb{R}^n$  to have an inverse, we must have  $m = n$ , and its coefficient matrix  $A$  must be nonsingular.

The invertibility of linear transformations on infinite-dimensional function spaces is more subtle. Here is a familiar example from calculus.

**Example 7.16.** The Fundamental Theorem of Calculus says, roughly, that differentiation  $D[f] = f'$  and (indefinite) integration  $J[f] = g$ , where  $g(x) = \int_a^x f(y) dy$ , are “inverse” operations. More precisely, the derivative of the indefinite integral of  $f$  is equal to  $f$ , and hence

$$D[J[f]] = D[g] = g' = f, \quad \text{since} \quad g'(x) = \frac{d}{dx} \int_a^x f(y) dy = f(x).$$

In other words, the composition  $D \circ J = I_{C^0[a,b]}$  defines the identity operator on the function space  $C^0[a,b]$ . On the other hand, if we integrate the derivative of a continuously differentiable function  $f \in C^1[a,b]$ , we obtain  $J[D[f]] = J[f'] = h$ , where

$$h(x) = \int_a^x f'(y) dy = f(x) - f(a) \neq f(x) \quad \text{unless} \quad f(a) = 0.$$

Therefore, the composition is *not* the identity operator:  $J \circ D \neq I_{C^1[a,b]}$ . In other words, the differentiation operator  $D$  is a left inverse for the integration operator  $J$  but not a right inverse!

If we restrict  $D$  to the subspace  $V = \{ f \mid f(a) = 0 \} \subset C^1[a,b]$  consisting of all continuously differentiable functions that vanish at the left-hand endpoint, then  $J: C^0[a,b] \rightarrow V$ , and  $D: V \rightarrow C^0[a,b]$  are, by the preceding argument, inverse linear operators:  $D \circ J = I_{C^0[a,b]}$ , and  $J \circ D = I_V$ . Note that  $V \subsetneq C^1[a,b] \subsetneq C^0[a,b]$ . Thus, we discover the curious and disconcerting infinite-dimensional phenomenon that  $J$  defines a one-to-one, invertible, linear map from a vector space  $C^0[a,b]$  to a proper subspace  $V \subsetneq C^0[a,b]$ . This paradoxical situation *cannot* occur in finite dimensions. A linear map  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  can be invertible only when its image is the entire space — because it represents multiplication by a nonsingular square matrix.

Two vector spaces  $V, W$  are said to be *isomorphic*, written  $V \simeq W$ , if there exists an invertible linear function  $L: V \rightarrow W$ . For example, if  $V$  is finite-dimensional, then  $V \simeq W$  if and only if  $W$  has the same dimension as  $V$ . In particular, if  $V$  has dimension  $n$ , then  $V \simeq \mathbb{R}^n$ . One way to construct the required invertible linear map is to choose a basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $V$ , and map it to the standard basis of  $\mathbb{R}^n$ , so  $L[\mathbf{v}_k] = \mathbf{e}_k$  for  $k = 1, \dots, n$ . In general, given  $\mathbf{v} = x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n$ , then, by linearity,

$$\begin{aligned} L[\mathbf{v}] &= L[x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n] = x_1 L[\mathbf{v}_1] + \dots + x_n L[\mathbf{v}_n] \\ &= x_1 \mathbf{e}_1 + \dots + x_n \mathbf{e}_n = (x_1, x_2, \dots, x_n)^T = \mathbf{x}, \end{aligned}$$

and hence  $L$  maps  $\mathbf{v}$  to the column vector  $\mathbf{x} = \mathbb{R}^n$  whose entries are its coordinates with respect to the chosen basis. The inverse  $L^{-1}: \mathbb{R}^n \rightarrow V$  maps  $\mathbf{x} \in \mathbb{R}^n$  to the element  $L^{-1}[\mathbf{x}] = x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n \in V$ . As the above example makes clear, isomorphism of infinite-dimensional vector spaces is more subtle, and one often imposes additional restrictions on the allowable linear maps.

## Exercises

- 7.1.51. Determine which of the following linear functions  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  has an inverse, and, if so, describe it: (a) the scaling transformation that doubles the length of each vector; (b) clockwise rotation by  $45^\circ$ ; (c) reflection through the  $y$ -axis; (d) orthogonal projection onto the line  $y = x$ ; (e) the shearing transformation defined by the matrix  $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ .
- 7.1.52. For each of the linear functions in Exercise 7.1.51, write down its matrix representative, the matrix representative of its inverse, and verify that the matrices are mutual inverses.
- 7.1.53. Let  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  be the linear function such that  $L[\mathbf{e}_1] = (1, -1)^T$ ,  $L[\mathbf{e}_2] = (3, -2)^T$ . Find  $L^{-1}[\mathbf{e}_1]$  and  $L^{-1}[\mathbf{e}_2]$ .
- 7.1.54. Let  $L: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be the linear function such that  $L[\mathbf{e}_1] = (2, 1, -1)^T$ ,  $L[\mathbf{e}_2] = (1, 2, 1)^T$ ,  $L[\mathbf{e}_3] = (-1, 2, 2)^T$ . Find  $L^{-1}[\mathbf{e}_1]$ ,  $L^{-1}[\mathbf{e}_2]$ , and  $L^{-1}[\mathbf{e}_3]$ .
- ◇ 7.1.55. Prove that the inverse of a linear transformation is unique; i.e., given  $L$ , there is at most one linear transformation  $M$  that can satisfy (7.19).
- ◇ 7.1.56. Let  $L: V \rightarrow W$  be a linear function. Suppose  $M, N: W \rightarrow V$  are linear functions that satisfy  $L \circ M = I_V = N \circ L$ . Prove that  $M = N = L^{-1}$ . Thus, a linear function may have only a left or a right inverse, but if it has both, then they must be the same.
- 7.1.57. Give an example of a matrix with a left inverse, but not a right inverse. Is your left inverse unique?
- ◇ 7.1.58. Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n$  is a basis for  $V$  and  $\mathbf{w}_1, \dots, \mathbf{w}_n$  a basis for  $W$ . (a) Prove that there is a unique linear function  $L: V \rightarrow W$  such that  $L[\mathbf{v}_i] = \mathbf{w}_i$  for  $i = 1, \dots, n$ . (b) Prove that  $L$  is invertible. (c) If  $V = W = \mathbb{R}^n$ , find a formula for the matrix representative of the linear functions  $L$  and  $L^{-1}$ . (d) Apply your construction to produce a linear function that takes: (i)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$  to  $\mathbf{w}_1 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ ,  $\mathbf{w}_2 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$ , (ii)  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$  to  $\mathbf{w}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ,  $\mathbf{w}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ , (iii)  $\mathbf{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$  to  $\mathbf{w}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ ,  $\mathbf{w}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{w}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ .
- 7.1.59. Suppose  $V, W \subset \mathbb{R}^n$  are subspaces of the same dimension. Prove that there is an invertible linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  that takes  $V$  to  $W$ . Hint: Use Exercise 7.1.58.
- ◇ 7.1.60. Let  $W, Z$  be complementary subspaces of a vector space  $V$ , as in Exercise 2.2.24. Let  $V/W$  denote the quotient vector space, as defined in Exercise 2.2.29. Show that the map  $L: Z \rightarrow V/W$  that maps  $L[\mathbf{z}] = [\mathbf{z}]_W$  defines an invertible linear map, and hence  $Z \simeq V/W$  are isomorphic vector spaces.
- ◇ 7.1.61. Let  $L: V \rightarrow W$  be a linear map. (a) Suppose  $V, W$  are finite-dimensional vector spaces, and let  $A$  be a matrix representative of  $L$ . Explain why we can identify  $\text{coker } A \simeq W/\text{img } A$  and  $\text{coimg } A = V/\ker A$  as quotient vector spaces, cf. Exercise 2.2.29.
- Remark.** These characterizations are used to give intrinsic definitions of the cokernel and coimage of a general linear function  $L: V \rightarrow W$  without any reference to a transpose (or, as defined below, adjoint) operation. Namely, set  $\text{coker } L \simeq W/\text{img } L$  and  $\text{coimg } L = V/\ker L$ . (b) The *index* of the linear map is defined as  $\text{index } L = \dim \ker L - \dim \text{coker } L$ , using the above intrinsic definitions. Prove that, when  $V, W$  are finite-dimensional,  $\text{index } L = \dim V - \dim W$ .

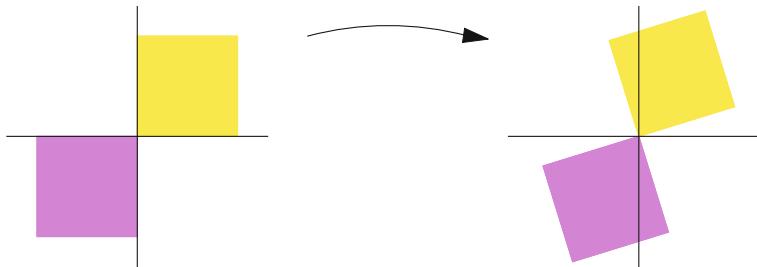


Figure 7.4. Rotation.

◇ 7.1.62. Let  $V$  be a finite-dimensional real inner product space and let  $V^*$  be its dual. Using Theorem 7.10, prove that the map  $J: V^* \rightarrow V$  that takes the linear function  $\ell \in V^*$  to the vector  $J[\ell] = \mathbf{a} \in V$  satisfying  $\ell[\mathbf{v}] = \langle \mathbf{a}, \mathbf{v} \rangle$  defines a linear isomorphism between the inner product space and its dual:  $V^* \simeq V$ .

7.1.63. (a) Prove that  $L[p] = p' + p$  defines an invertible linear map on the space  $\mathcal{P}^{(2)}$  of quadratic polynomials. Find a formula for its inverse.

(b) Does the derivative  $D[p] = p'$  have either a left or a right inverse on  $\mathcal{P}^{(2)}$ ?

◇ 7.1.64. (a) Show that the set of all functions of the form  $f(x) = (ax^2 + bx + c)e^x$  for  $a, b, c \in \mathbb{R}$  is a vector space. What is its dimension? (b) Show that the derivative  $D[f(x)] = f'(x)$  defines an invertible linear transformation on this vector space, and determine its inverse. (c) Generalize your result in part (b) to the infinite-dimensional vector space consisting of all functions of the form  $p(x)e^x$ , where  $p(x)$  is an arbitrary polynomial.

## 7.2 Linear Transformations

Consider a linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  that maps  $n$ -dimensional Euclidean space to itself. The function  $L$  maps a point  $\mathbf{x} \in \mathbb{R}^n$  to its image point  $L[\mathbf{x}] = A\mathbf{x}$ , where  $A$  is its  $n \times n$  matrix representative. As such, it can be assigned a geometrical interpretation that leads to further insight into the nature and scope of linear functions on Euclidean space. The geometrically inspired term *linear transformation* is often used to refer to such linear functions. The two-, three-, and four-dimensional (viewing time as the fourth dimension of space-time) cases have particular relevance to our physical universe. Many of the notable maps that appear in geometry, computer graphics, elasticity, symmetry, crystallography, and Einstein's special relativity, to name a few, are defined by linear transformations.

Most of the important classes of linear transformations already appear in the two-dimensional case. Every linear function  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  has the form

$$L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix}, \quad \text{where} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad (7.20)$$

is an arbitrary  $2 \times 2$  matrix. We have already encountered the *rotation matrices*

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}, \quad (7.21)$$

whose effect is to rotate every vector in  $\mathbb{R}^2$  through an angle  $\theta$ ; in Figure 7.4 we illustrate the effect on a couple of square regions in the plane. Planar rotations coincide with  $2 \times 2$

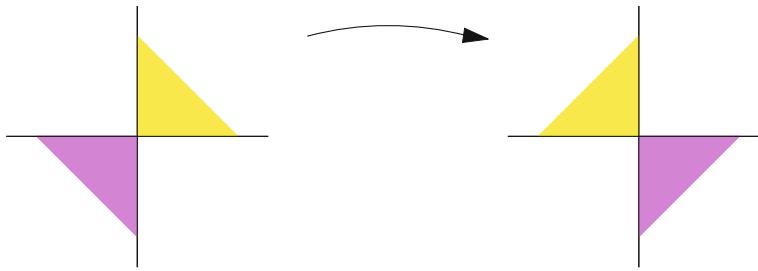


Figure 7.5. Reflection through the  $y$ -axis.

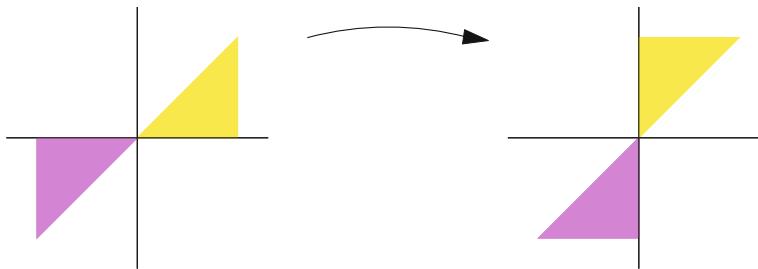


Figure 7.6. Reflection through the Diagonal.

proper orthogonal matrices, meaning matrices  $Q$  that satisfy

$$Q^T Q = I, \quad \det Q = +1. \quad (7.22)$$

The improper orthogonal matrices, i.e., those with determinant  $-1$ , define *reflections*. For example, the matrix

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \text{ corresponds to the linear transformation } L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix}, \quad (7.23)$$

which reflects the plane through the  $y$ -axis. It can be visualized by thinking of the  $y$ -axis as a mirror, as illustrated in Figure 7.5. Another simple example is the improper orthogonal matrix

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad \text{The corresponding linear transformation } L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x \end{pmatrix} \quad (7.24)$$

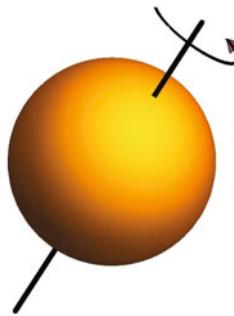
is a reflection through the diagonal line  $y = x$ , as illustrated in Figure 7.6.

A similar classification of orthogonal matrices carries over to three-dimensional (and even higher-dimensional) space. The proper orthogonal matrices correspond to rotations and the improper orthogonal matrices to reflections, or, more generally, reflections combined with rotations. For example, the proper orthogonal matrix

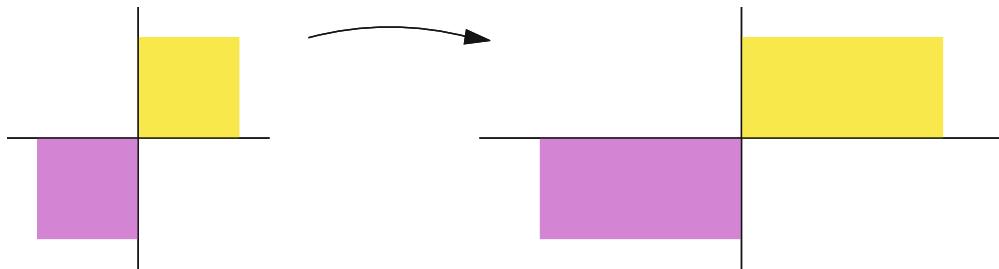
$$Z_\theta = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (7.25)$$

corresponds to a counterclockwise rotation through an angle  $\theta$  around the  $z$ -axis, while

$$Y_\varphi = \begin{pmatrix} \cos \varphi & 0 & -\sin \varphi \\ 0 & 1 & 0 \\ \sin \varphi & 0 & \cos \varphi \end{pmatrix} \quad (7.26)$$



**Figure 7.7.** A Three-Dimensional Rotation.



**Figure 7.8.** Stretch along the  $x$ -axis.

corresponds to a clockwise rotation through an angle  $\varphi$  around the  $y$ -axis. In general, a proper orthogonal matrix  $Q = (\mathbf{u}_1 \ \mathbf{u}_2 \ \mathbf{u}_3)$  with columns  $\mathbf{u}_i = Q\mathbf{e}_i$  corresponds to the rotation in which the standard basis vectors  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  are rotated to new positions given by the orthonormal basis  $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ . It can be shown — see Exercise 8.2.44 — that *every*  $3 \times 3$  orthogonal matrix corresponds to a rotation around a line through the origin in  $\mathbb{R}^3$  — the axis of the rotation, as sketched in [Figure 7.7](#).

Since the product of two (proper) orthogonal matrices is also (proper) orthogonal, the composition of two rotations is also a rotation. Unlike the planar case, the order in which the rotations are performed is important! Multiplication of  $n \times n$  orthogonal matrices is *not* commutative when  $n \geq 3$ . For example, rotating first around the  $z$ -axis and then rotating around the  $y$ -axis does *not* have the same effect as first rotating around the  $y$ -axis and then around the  $z$ -axis. If you don't believe this, try it out with a solid object such as this book. Rotate through  $90^\circ$ , say, around each axis; the final configuration of the book will depend upon the order in which you do the rotations. Then prove this mathematically by showing that the two rotation matrices (7.25, 26) do not commute.

Other important linear transformations arise from elementary matrices. First, the elementary matrices corresponding to the third type of row operations — multiplying a row by a scalar — correspond to simple stretching transformations. For example, if

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{then the linear transformation} \quad L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x \\ y \end{pmatrix}$$

has the effect of stretching along the  $x$ -axis by a factor of 2; see [Figure 7.8](#). A negative diagonal entry corresponds to a reflection followed by a stretch. For example, the elementary

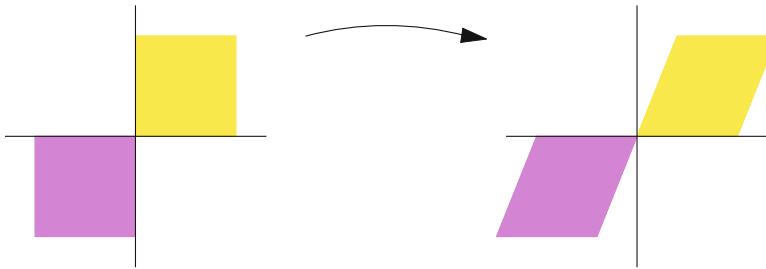


Figure 7.9. Shear in the  $x$  Direction.

matrix

$$\begin{pmatrix} -2 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

corresponds to a reflection through the  $y$ -axis followed by a stretch along the  $x$ -axis. In this case, the order of these operations is immaterial, since the matrices commute.

In the  $2 \times 2$  case, there is only one type of elementary row interchange, namely the matrix  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ , which corresponds to a reflection through the diagonal  $y = x$ , as in (7.24).

The elementary matrices of type #1 correspond to *shearing transformations* of the plane. For example, the matrix

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{represents the linear transformation} \quad L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 2y \\ y \end{pmatrix},$$

which has the effect of shearing the plane along the  $x$ -axis. The constant 2 will be called the *shear factor*, and can be either positive or negative. Under the shearing transformation, each point moves parallel to the  $x$ -axis by an amount proportional to its (signed) distance from the axis. Similarly, the elementary matrix

$$\begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \quad \text{represents the linear transformation} \quad L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y - 3x \end{pmatrix},$$

which is a shear along the  $y$ -axis of magnitude  $-3$ . As illustrated in Figure 7.9, shears map rectangles to parallelograms; distances are altered, but areas are unchanged.

All of the preceding linear maps are invertible, and so are represented by nonsingular matrices. Besides the zero map/matrix, which sends every point  $\mathbf{x} \in \mathbb{R}^2$  to the origin, the simplest singular map is

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \text{corresponding to the linear transformation} \quad L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix},$$

which defines the orthogonal projection of the vector  $(x, y)^T$  onto the  $x$ -axis. Other rank one matrices represent various kinds of projections from the plane to a line through the origin; see Exercise 7.2.16 for details.

A similar classification of linear maps can be established in higher dimensions. The linear transformations constructed from elementary matrices can be built up from the

following four basic types:

- (i) a stretch in a single coordinate direction;
- (ii) a reflection through a coordinate plane;<sup>†</sup>
- (iii) a reflection through a diagonal plane;
- (iv) a shear along a coordinate axis.

Moreover, we already proved — see (1.47) — that every nonsingular matrix can be written as a product of elementary matrices. This has the remarkable consequence that *every* invertible linear transformation can be constructed from a sequence of elementary stretches, reflections, and shears. In addition, there is one further, non-invertible, type of basic linear transformation:

- (v) an orthogonal projection onto a lower-dimensional subspace.

All linear transformations of  $\mathbb{R}^n$  can be built up, albeit non-uniquely, as a composition of these five basic types.

**Example 7.17.** Consider the matrix  $A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}$  corresponding to a plane rotation through  $\theta = 30^\circ$ , cf. (7.21). Rotations are not elementary linear transformations. To express this particular rotation as a product of elementary matrices, we need to perform the Gauss-Jordan Elimination procedure to reduce it to the identity matrix. Let us indicate the basic steps:

$$\begin{array}{ll} E_1 = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\sqrt{3}} & 1 \end{pmatrix}, & E_1 A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix}, \\ E_2 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix}, & E_2 E_1 A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ 0 & 1 \end{pmatrix}, \\ E_3 = \begin{pmatrix} \frac{2}{\sqrt{3}} & 0 \\ 0 & 1 \end{pmatrix}, & E_3 E_2 E_1 A = \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix}, \\ E_4 = \begin{pmatrix} 1 & \frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix}, & E_4 E_3 E_2 E_1 A = I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{array}$$

We conclude that

$$\begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} = A = E_1^{-1} E_2^{-1} E_3^{-1} E_4^{-1} = \begin{pmatrix} 1 & 0 \\ \frac{1}{\sqrt{3}} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix}.$$

As a result, a  $30^\circ$  rotation can be effected by composing the following elementary transformations in the prescribed order, bearing in mind that the last matrix in the product will act first on the vector  $\mathbf{x}$ :

- (1) First, a shear in the  $x$  direction with shear factor  $-\frac{1}{\sqrt{3}}$ .
- (2) Then a stretch (or, rather, a contraction) in the direction of the  $x$ -axis by a factor of  $\frac{\sqrt{3}}{2}$ .
- (3) Then a stretch in the  $y$  direction by the reciprocal factor  $\frac{2}{\sqrt{3}}$ .
- (4) Finally, a shear in the direction of the  $y$ -axis with shear factor  $\frac{1}{\sqrt{3}}$ .

---

<sup>†</sup> In  $n$ -dimensional space, this should read “hyperplane”, i.e., a subspace of dimension  $n - 1$ .

The fact that this combination of elementary transformations results in a pure rotation is surprising and non-obvious.

## Exercises

7.2.1. For each of the following linear transformations  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ , find a matrix representative, and then describe its effect on (i) the  $x$ -axis; (ii) the unit square  $S = \{0 \leq x, y \leq 1\}$ ; (iii) the unit disk  $D = \{x^2 + y^2 \leq 1\}$ : (a) counterclockwise rotation by  $45^\circ$ ; (b) rotation by  $180^\circ$ ; (c) reflection in the line  $y = 2x$ ; (d) shear along the  $y$ -axis of magnitude 2; (e) shear along the line  $x = y$  of magnitude 3; (f) orthogonal projection on the line  $y = 2x$ .

7.2.2. Let  $L$  be the linear transformation represented by the matrix  $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ . Show that  $L^2 = L \circ L$  is rotation by  $180^\circ$ . Is  $L$  itself a rotation or a reflection?

7.2.3. Let  $L$  be the linear transformation determined by  $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . Show  $L^2 = I$ , and interpret geometrically.

7.2.4. What is the geometric interpretation of the linear transformation with matrix

$A = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix}$ ? Use this to explain why  $A^2 = I$ .

7.2.5. Describe the image of the line  $\ell$  that goes through the points  $\begin{pmatrix} -2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \end{pmatrix}$  under the linear transformation  $\begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix}$ .

7.2.6. Draw the parallelogram spanned by the vectors  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$  and  $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ . Then draw its image under the linear transformations defined by the following matrices:

$$(a) \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix}, \quad (d) \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},$$

$$(e) \begin{pmatrix} -1 & -2 \\ 2 & 1 \end{pmatrix}, \quad (f) \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad (g) \begin{pmatrix} 2 & -1 \\ -4 & 2 \end{pmatrix}.$$

7.2.7. Find a linear transformation that maps the unit circle  $x^2 + y^2 = 1$  to the ellipse  $\frac{1}{4}x^2 + \frac{1}{9}y^2 = 1$ . Is your answer unique?

7.2.8. Find a linear transformation that maps the unit sphere  $x^2 + y^2 + z^2 = 1$  to the ellipsoid  $x^2 + \frac{1}{4}y^2 + \frac{1}{16}z^2 = 1$ .

7.2.9. *True or false:* A linear transformation  $L: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  maps

- (a) straight lines to straight lines; (b) triangles to triangles; (c) squares to squares;
- (d) circles to circles; (e) ellipses to ellipses.

◇ 7.2.10. (a) Prove that the linear transformation associated with the improper orthogonal matrix  $\begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}$  is a reflection through the line that makes an angle  $\frac{1}{2}\theta$  with the  $x$ -axis.  
 (b) Show that the composition of two such reflections, with angles  $\theta, \varphi$ , is a rotation. What is the angle of the rotation? Does the composition depend upon the order of the two reflections?

7.2.11. (a) Find the matrix in  $\mathbb{R}^3$  that corresponds to a counterclockwise rotation around the  $x$ -axis through an angle  $60^\circ$ . (b) Write it as a product of elementary matrices, and interpret each of the factors.

◇ 7.2.12. Let  $L \subset \mathbb{R}^2$  be the line through the origin in the direction of a unit vector  $\mathbf{u}$ . (a) Prove that the matrix representative of reflection through  $L$  is  $R = 2\mathbf{u}\mathbf{u}^T - \mathbf{I}$ . (b) Find the corresponding formula for reflection through the line in the direction of a general nonzero vector  $\mathbf{v} \neq \mathbf{0}$ . (c) Determine the matrix representative for reflection through the line in the direction (i)  $(1, 0)^T$ , (ii)  $(\frac{3}{5}, -\frac{4}{5})^T$ , (iii)  $(1, 1)^T$ , (iv)  $(2, -3)^T$ .

7.2.13. Decompose the following matrices into a product of elementary matrices. Then interpret each of the factors as a linear transformation.

$$(a) \begin{pmatrix} 0 & 2 \\ -3 & 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 2 & 1 & 1 \end{pmatrix}.$$

7.2.14. (a) Prove that  $\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix} \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ , where  $a = -\tan \frac{1}{2}\theta$  and  $b = \sin \theta$ . (b) Is the factorization valid for all values of  $\theta$ ? (c) Interpret the factorization geometrically. **Remark.** The factored version is less prone to numerical errors due to round-off, and so can be used when extremely accurate numerical computations involving rotations are required.

7.2.15. Determine the matrix representative for orthogonal projection  $P: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  on the line through the origin in the direction (a)  $(1, 0)^T$ , (b)  $(1, 1)^T$ , (c)  $(2, -3)^T$ .

◇ 7.2.16. (a) Prove that every  $2 \times 2$  matrix of rank 1 can be written in the form  $A = \mathbf{u}\mathbf{v}^T$  where  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$  are non-zero column vectors. (b) Which rank one matrices correspond to orthogonal projection onto a one-dimensional subspace of  $\mathbb{R}^2$ ?

7.2.17. Give a geometrical interpretation of the linear transformations on  $\mathbb{R}^3$  defined by each of the six  $3 \times 3$  permutation matrices (1.30).

7.2.18. Write down the  $3 \times 3$  matrix  $X_\psi$  representing a clockwise rotation in  $\mathbb{R}^3$  around the  $x$ -axis by angle  $\psi$ .

7.2.19. Explain why the linear map defined by  $-\mathbf{I}$  defines a rotation in two-dimensional space, but a reflection in three-dimensional space.

◇ 7.2.20. Let  $\mathbf{u} = (u_1, u_2, u_3)^T \in \mathbb{R}^3$  be a unit vector. Show that  $Q_\pi = 2\mathbf{u}\mathbf{u}^T - \mathbf{I}$  represents rotation around the axis  $\mathbf{u}$  through an angle  $\pi$ .

◇ 7.2.21. Let  $\mathbf{u} \in \mathbb{R}^3$  be a unit vector. (a) Explain why the *elementary reflection matrix*  $R = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T$  represents a reflection through the plane orthogonal to  $\mathbf{u}$ . (b) Prove that  $R$  is an orthogonal matrix. Is it proper or improper? (c) Write out  $R$  when  $\mathbf{u} = (i) \left(\frac{3}{5}, 0, -\frac{4}{5}\right)^T$ , (ii)  $\left(\frac{3}{13}, \frac{4}{13}, -\frac{12}{13}\right)^T$ , (iii)  $\left(\frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)^T$ . (d) Give a geometrical explanation why  $Q_\pi = -R$  represents the rotation of Exercise 7.2.20.

◇ 7.2.22. Let  $\mathbf{a} \in \mathbb{R}^3$ , and let  $Q$  be any  $3 \times 3$  rotation matrix such that  $Q\mathbf{a} = \mathbf{e}_3$ . (a) Show, using the notation of (7.25), that  $R_\theta = Q^T Z_\theta Q$  represents rotation around  $\mathbf{a}$  by angle  $\theta$ . (b) Verify this formula in the case  $\mathbf{a} = \mathbf{e}_2$  by comparing with (7.26).

◇ 7.2.23. *Quaternions:* The *skew field*  $\mathbb{H}$  of quaternions can be identified with the vector space  $\mathbb{R}^4$  equipped with a *noncommutative* multiplication operation. The standard basis vectors  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$  are traditionally denoted by the letters  $1, i, j, k$ ; the vector  $(a, b, c, d)^T \in \mathbb{R}^4$  corresponds to the quaternion  $q = a + bi + cj + dk$ . Quaternion addition coincides with vector addition. Quaternion multiplication is defined so that

$$1q = q = q1, \quad i^2 = j^2 = k^2 = -1, \quad ij = k = -ji, \quad ik = -j = -ki, \quad jk = i = -kj,$$

along with the distributive laws

$$(q+r)s = qs + rs, \quad q(r+s) = qr + qs, \quad \text{for all } q, r, s \in \mathbb{H}.$$

(a) Compute the following quaternion products: (i)  $j(2-3j+k)$ , (ii)  $(1+i)(1-2i+j)$ , (iii)  $(1+i-j-3k)^2$ , (iv)  $(2+2i+3j-k)(2-2i-3j+k)$ . (b) The *conjugate* of the quaternion  $q = a + bi + cj + dk$  is defined to be  $\bar{q} = a - bi - cj - dk$ . Prove that  $q\bar{q} = \|q\|^2 = \bar{q}q$ , where  $\|\cdot\|$  is the usual Euclidean norm on  $\mathbb{R}^4$ . (c) Prove that quaternion multiplication is associative. (d) Let  $q = a + bi + cj + dk \in \mathbb{H}$ . Show that  $L_q[r] = qr$  and  $R_q[r] = rq$  define linear transformations on the vector space  $\mathbb{H} \simeq \mathbb{R}^4$ . Write down their  $4 \times 4$  matrix representatives, and observe that they are not the same, since quaternion multiplication is not commutative. (e) Show that  $L_q$  and  $R_q$  are orthogonal matrices if  $\|q\|^2 = a^2 + b^2 + c^2 + d^2 = 1$ . (f) We can identify a quaternion  $q = bi + cj + dk$  with zero real part,  $a = 0$ , with a vector  $\mathbf{q} = (b, c, d)^T \in \mathbb{R}^3$ . Show that, in this case, the quaternion product  $qr = \mathbf{q} \times \mathbf{r} - \mathbf{q} \cdot \mathbf{r}$  can be identified with the difference between the cross and dot product of the two vectors. Which vector identities result from the associativity of quaternion multiplication? **Remark.** The *quaternions* were discovered by the Irish mathematician William Rowan Hamilton in 1843. Much of our modern vector calculus notation is of quaternionic origin, [17].

## Change of Basis

Sometimes a linear transformation represents an elementary geometrical transformation, but this is not evident because the matrix happens to be written in the “wrong” coordinates. The characterization of linear functions from  $\mathbb{R}^n$  to  $\mathbb{R}^m$  as multiplication by  $m \times n$  matrices in Theorem 7.5 relies on using the standard bases for both the domain and codomain. In many cases, these bases are not particularly well adapted to the linear transformation in question, and one can often gain additional insight by adopting more suitable bases. To this end, we first need to understand how to rewrite a linear transformation in terms of a new basis.

The following result says that, in *any* basis, a linear function on finite-dimensional vector spaces can always be realized by matrix multiplication of the coordinates. But bear in mind that the particular matrix representative will depend upon the choice of bases.

**Theorem 7.18.** Let  $L: V \rightarrow W$  be a linear function. Suppose  $V$  has basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  and  $W$  has basis  $\mathbf{w}_1, \dots, \mathbf{w}_m$ . We can write

$$\mathbf{v} = x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n \in V, \quad \mathbf{w} = y_1 \mathbf{w}_1 + \dots + y_m \mathbf{w}_m \in W,$$

where  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  are the coordinates of  $\mathbf{v}$  relative to the basis of  $V$  and  $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$  are those of  $\mathbf{w}$  relative to the basis of  $W$ . Then, in these coordinates, the linear function  $\mathbf{w} = L[\mathbf{v}]$  is given by multiplication by an  $m \times n$  matrix  $B$ , so  $\mathbf{y} = B\mathbf{x}$ .

*Proof:* We mimic the proof of Theorem 7.5, replacing the standard basis vectors by more general basis vectors. In other words, we will apply  $L$  to the basis vectors of  $V$  and express the result as a linear combination of the basis vectors in  $W$ . Specifically, we write

$$L[\mathbf{v}_j] = \sum_{i=1}^m b_{ij} \mathbf{w}_i.$$

The coefficients  $b_{ij}$  form the entries of the desired coefficient matrix. Indeed, by linearity,

$$L[\mathbf{v}] = L[x_1 \mathbf{v}_1 + \dots + x_n \mathbf{v}_n] = x_1 L[\mathbf{v}_1] + \dots + x_n L[\mathbf{v}_n] = \sum_{i=1}^m \left( \sum_{j=1}^n b_{ij} x_j \right) \mathbf{w}_i,$$

and so  $y_i = \sum_{j=1}^n b_{ij} x_j$ , as claimed.

*Q.E.D.*

Suppose that the linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is represented by a certain  $m \times n$  matrix  $A$  relative to the standard bases  $\mathbf{e}_1, \dots, \mathbf{e}_n$  and  $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_m$  of the domain and codomain. If we introduce alternative bases for  $\mathbb{R}^n$  and  $\mathbb{R}^m$ , then the *same* linear transformation may have a completely different matrix representation. Therefore, different matrices may represent the same underlying linear transformation, but relative to different bases of its domain and codomain.

**Example 7.19.** Consider the linear transformation

$$L[\mathbf{x}] = L \left[ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right] = \begin{pmatrix} x_1 - x_2 \\ 2x_1 + 4x_2 \end{pmatrix}, \quad (7.27)$$

which we write in the standard, Cartesian coordinates on  $\mathbb{R}^2$ . The corresponding coefficient matrix

$$A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix} \quad (7.28)$$

is the matrix representation of  $L$ , relative to the standard basis  $\mathbf{e}_1, \mathbf{e}_2$  of  $\mathbb{R}^2$ , and can be read directly off the explicit formula (7.27):

$$L[\mathbf{e}_1] = L \left[ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right] = \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \mathbf{e}_1 + 2 \mathbf{e}_2, \quad L[\mathbf{e}_2] = L \left[ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} -1 \\ 4 \end{pmatrix} = -\mathbf{e}_1 + 4\mathbf{e}_2.$$

Let us see what happens if we replace the standard basis by the alternative basis

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

What is the corresponding matrix formulation of the same linear transformation? According to the recipe of Theorem 7.18, we must compute

$$L[\mathbf{v}_1] = \begin{pmatrix} 2 \\ -2 \end{pmatrix} = 2\mathbf{v}_1, \quad L[\mathbf{v}_2] = \begin{pmatrix} 3 \\ -6 \end{pmatrix} = 3\mathbf{v}_2.$$

The linear transformation acts by stretching in the direction  $\mathbf{v}_1$  by a factor of 2 and simultaneously stretching in the direction  $\mathbf{v}_2$  by a factor of 3. Therefore, the matrix form of  $L$  with respect to this new basis is the diagonal matrix

$$D = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}. \quad (7.29)$$

In general,

$$L[a\mathbf{v}_1 + b\mathbf{v}_2] = 2a\mathbf{v}_1 + 3b\mathbf{v}_2,$$

and the effect is to multiply the new basis coordinates  $\mathbf{a} = (a, b)^T$  by the diagonal matrix  $D$ . Both (7.28) and (7.29) represent the *same* linear transformation — the former in the standard basis and the latter in the new basis. The hidden geometry of this linear transformation is thereby exposed through an inspired choice of basis. The secret behind such well-adapted bases will be revealed in Chapter 8.

How does one effect a change of basis in general? According to formula (2.23), if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis of  $\mathbb{R}^n$ , then the coordinates  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$  of a vector

$$(x_1, x_2, \dots, x_n)^T = \mathbf{x} = y_1\mathbf{v}_1 + y_2\mathbf{v}_2 + \dots + y_n\mathbf{v}_n$$

are found by solving the linear system

$$S \mathbf{y} = \mathbf{x}, \quad \text{where} \quad S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n) \quad (7.30)$$

is the nonsingular  $n \times n$  matrix whose columns are the basis vectors.

Consider first a linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  from  $\mathbb{R}^n$  to itself. When written in terms of the standard basis,  $L[\mathbf{x}] = A\mathbf{x}$  has a certain  $n \times n$  coefficient matrix  $A$ . To change to the new basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ , we use (7.30) to rewrite the standard  $\mathbf{x}$  coordinates in terms of the new  $\mathbf{y}$  coordinates. We also need to find the coordinates  $\mathbf{g}$  of an image vector  $\mathbf{f} = A\mathbf{x}$  with respect to the new basis. By the same reasoning that led to (7.30), its new coordinates are found by solving the linear system  $\mathbf{f} = S\mathbf{g}$ . Therefore, the new codomain coordinates are expressed in terms of the new domain coordinates via

$$\mathbf{g} = S^{-1}\mathbf{f} = S^{-1}A\mathbf{x} = S^{-1}AS\mathbf{y} = B\mathbf{y}.$$

We conclude that, in the new basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ , the matrix form of our linear transformation is

$$B = S^{-1}AS, \quad \text{where} \quad S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n). \quad (7.31)$$

Two matrices  $A$  and  $B$  that are related by such an equation for some nonsingular matrix  $S$  are called *similar*. Similar matrices represent the *same* linear transformation, but relative to *different* bases of the underlying vector space  $\mathbb{R}^n$ , the matrix  $S$  serving to encode the *change of basis*.

**Example 7.19 (continued).** Returning to the preceding example, we assemble the new basis vectors to form the change of basis matrix  $S = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$ , and verify that

$$S^{-1}AS = \begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} = D,$$

reconfirming our earlier computation.

More generally, a linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is represented by an  $m \times n$  matrix  $A$  with respect to the standard bases on both the domain and codomain. What happens if we introduce a new basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  on the domain space  $\mathbb{R}^n$  and a new basis  $\mathbf{w}_1, \dots, \mathbf{w}_m$  on the codomain  $\mathbb{R}^m$ ? Arguing as above, we conclude that the matrix representative of  $L$  with respect to these new bases is given by

$$B = T^{-1}AS, \quad (7.32)$$

where  $S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  is the domain basis matrix, while  $T = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_m)$  is the image basis matrix.

In particular, suppose that  $L$  has rank  $r = \dim \text{img } A = \dim \text{coimg } A$ . Let us choose a basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $\mathbb{R}^n$  such that  $\mathbf{v}_1, \dots, \mathbf{v}_r$  form a basis of  $\text{coimg } A$ , while  $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$  form a basis for  $\ker A = (\text{coimg } A)^\perp$ . According to Theorem 4.49, the image vectors  $\mathbf{w}_1 = L[\mathbf{v}_1], \dots, \mathbf{w}_r = L[\mathbf{v}_r]$  form a basis for  $\text{img } A$ , while  $L[\mathbf{v}_{r+1}] = \dots = L[\mathbf{v}_n] = \mathbf{0}$ . We further choose a basis  $\mathbf{w}_{r+1}, \dots, \mathbf{w}_m$  for  $\text{coker } A = (\text{img } A)^\perp$ , and note that the combination  $\mathbf{w}_1, \dots, \mathbf{w}_m$  is a basis for  $\mathbb{R}^m$ . The matrix form of  $L$  relative to these two adapted bases is

simply

$$B = T^{-1}AS = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}. \quad (7.33)$$

In this matrix, the first  $r$  columns have a single 1 in the diagonal slot, indicating that the first  $r$  basis vectors of the domain space are mapped to the first  $r$  basis vectors of the codomain, while the last  $n - r$  columns are all zero, indicating that the last  $n - r$  basis vectors in the domain are all mapped to  $\mathbf{0}$ . Thus, by a suitable choice of bases on both the domain and codomain, every linear transformation has an extremely simple *canonical form* (7.33) that depends *only* on its rank.

**Example 7.20.** According to the example following Theorem 2.49, the matrix

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

has rank 2. Based on those calculations, we choose the domain space basis

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 0 \\ -2 \\ 4 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} -2 \\ 0 \\ 2 \\ 1 \end{pmatrix},$$

noting that  $\mathbf{v}_1, \mathbf{v}_2$  are a basis for  $\text{coimg } A$ , while  $\mathbf{v}_3, \mathbf{v}_4$  are a basis for  $\ker A$ . For our basis of the codomain, we first compute  $\mathbf{w}_1 = A\mathbf{v}_1$  and  $\mathbf{w}_2 = A\mathbf{v}_2$ , which form a basis for  $\text{img } A$ . We supplement these by the single basis vector  $\mathbf{w}_3$  for  $\text{coker } A$ , where

$$\mathbf{w}_1 = \begin{pmatrix} 10 \\ -34 \\ 17 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 6 \\ -4 \\ 2 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 1 \end{pmatrix}.$$

By construction,  $B[\mathbf{v}_1] = \mathbf{w}_1$ ,  $B[\mathbf{v}_2] = \mathbf{w}_2$ ,  $B[\mathbf{v}_3] = B[\mathbf{v}_4] = \mathbf{0}$ , and thus the canonical matrix form of this particular linear function is given in terms of these two bases as

$$B = T^{-1}AS = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where the bases are assembled to form the matrices

$$S = \begin{pmatrix} 2 & 0 & \frac{1}{2} & -2 \\ -1 & 0 & 1 & 0 \\ 1 & -2 & 0 & 2 \\ 2 & 4 & 0 & 1 \end{pmatrix}, \quad T = \begin{pmatrix} 10 & 6 & 0 \\ -34 & -4 & \frac{1}{2} \\ 17 & 2 & 1 \end{pmatrix}.$$

## Exercises

7.2.24. Find the matrix form of the linear transformation  $L(x, y) = \begin{pmatrix} x - 4y \\ -2x + 3y \end{pmatrix}$  with respect to the following bases of  $\mathbb{R}^2$ :

$$(a) \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, (b) \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix}, (c) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, (d) \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, (e) \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

7.2.25. Find the matrix form of  $L[\mathbf{x}] = \begin{pmatrix} -3 & 2 & 2 \\ -3 & 1 & 3 \\ -1 & 2 & 0 \end{pmatrix} \mathbf{x}$  with respect to the following bases of  $\mathbb{R}^3$ :

$$(a) \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix}, (b) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, (c) \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -2 \\ 1 \end{pmatrix}.$$

7.2.26. Find bases of the domain and codomain that place the following matrices in the

canonical form (7.33). Use (7.32) to check your answer. (a)  $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ ,  
(b)  $\begin{pmatrix} 1 & -3 & 4 \\ -2 & 6 & -8 \end{pmatrix}$ , (c)  $\begin{pmatrix} 2 & 3 \\ 0 & 4 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & -1 \\ 2 & 1 & 0 \end{pmatrix}$ , (e)  $\begin{pmatrix} 1 & 3 & 0 & 1 \\ 2 & 6 & 1 & -2 \\ -1 & -3 & -1 & 3 \\ 0 & 0 & -1 & 4 \end{pmatrix}$ .

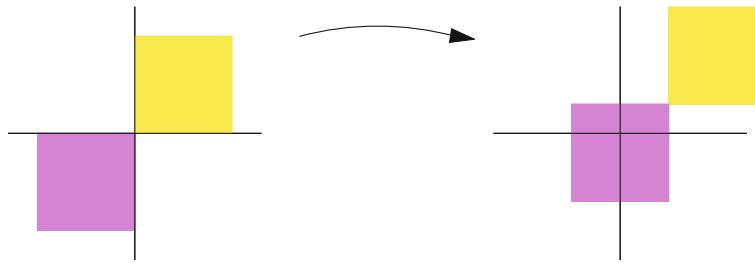
7.2.27. (a) Show that every invertible linear function  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  can be represented by the identity matrix by choosing appropriate (and not necessarily the same) bases on the domain and codomain. (b) Which linear transformations are represented by the identity matrix when the domain and codomain are required to have the same basis? (c) Find bases of  $\mathbb{R}^2$  so that the following linear transformations are represented by the identity matrix: (i) the scaling map  $S[\mathbf{x}] = 2\mathbf{x}$ ; (ii) counterclockwise rotation by  $45^\circ$ ; (iii) the shear  $\begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$ .

◇ 7.2.28. Suppose a linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is represented by a symmetric matrix with respect to the standard basis  $\mathbf{e}_1, \dots, \mathbf{e}_n$ . (a) Prove that its matrix representative with respect to any orthonormal basis  $\mathbf{u}_1, \dots, \mathbf{u}_n$  is symmetric. (b) Is it symmetric when expressed in terms of a non-orthonormal basis?

◇ 7.2.29. In this exercise, we show that every inner product  $\langle \cdot, \cdot \rangle$  on  $\mathbb{R}^n$  can be reduced to the dot product when expressed in a suitably adapted basis. (a) Specifically, prove that there exists a basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $\mathbb{R}^n$  such that  $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n c_i d_i = \mathbf{c} \cdot \mathbf{d}$ , where  $\mathbf{c} = (c_1, c_2, \dots, c_n)^T$  are the coordinates of  $\mathbf{x}$  and  $\mathbf{d} = (d_1, d_2, \dots, d_n)^T$  those of  $\mathbf{y}$  with respect to the basis. Is the basis uniquely determined? (b) Find bases that reduce the following inner products to the dot product on  $\mathbb{R}^2$ :

$$(i) \langle \mathbf{v}, \mathbf{w} \rangle = 2v_1w_1 + 3v_2w_2, \quad (ii) \langle \mathbf{v}, \mathbf{w} \rangle = v_1w_1 - v_1w_2 - v_2w_1 + 3v_2w_2.$$

◇ 7.2.30. *Dual functions:* Let  $L: V \rightarrow W$  be a linear function between vector spaces. The *dual linear function*, denoted by  $L^*: W^* \rightarrow V^*$  (note the change in direction) is defined so that  $L^*(m) = m \circ L$  for all linear functions  $m \in W^*$ . (a) Prove that  $L^*$  is a linear function. (b) If  $M: W \rightarrow Z$  is linear, prove that  $(M \circ L)^* = L^* \circ M^*$ . (c) Suppose  $\dim V = n$  and  $\dim W = m$ . Prove that if  $L$  is represented by the  $m \times n$  matrix  $A$  with respect to bases of  $V, W$ , then  $L^*$  is represented by the  $n \times m$  transposed matrix  $A^T$  with respect to the dual bases, as defined in Exercise 7.1.32.



**Figure 7.10.** Translation.

- ◇ 7.2.31. Suppose  $A$  is an  $m \times n$  matrix. (a) Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis of  $\mathbb{R}^n$ , and  $A\mathbf{v}_i = \mathbf{w}_i \in \mathbb{R}^m$ , for  $i = 1, \dots, n$ . Prove that the vectors  $\mathbf{v}_1, \dots, \mathbf{v}_n, \mathbf{w}_1, \dots, \mathbf{w}_n$ , serve to uniquely specify  $A$ .  
 (b) Write down a formula for  $A$ .

### 7.3 Affine Transformations and Isometries

Not every transformation of importance in geometrical applications arises as a linear function. A simple example is a *translation*, whereby all the points in  $\mathbb{R}^n$  are moved in the same direction by a common distance. The function that accomplishes this is

$$T[\mathbf{x}] = \mathbf{x} + \mathbf{b}, \quad \mathbf{x} \in \mathbb{R}^n, \quad (7.34)$$

where  $\mathbf{b} \in \mathbb{R}^n$  determines the direction and the distance that the points are translated. Except in the trivial case  $\mathbf{b} = \mathbf{0}$ , the translation  $T$  is *not* a linear function because

$$T[\mathbf{x} + \mathbf{y}] = \mathbf{x} + \mathbf{y} + \mathbf{b} \neq T[\mathbf{x}] + T[\mathbf{y}] = \mathbf{x} + \mathbf{y} + 2\mathbf{b}.$$

Or, even more simply, we note that  $T[\mathbf{0}] = \mathbf{b}$ , which must be  $\mathbf{0}$  if  $T$  is to be linear.

Combining translations and linear functions leads us to an important class of geometrical transformations.

**Definition 7.21.** A function  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$  of the form

$$F[\mathbf{x}] = A\mathbf{x} + \mathbf{b}, \quad (7.35)$$

where  $A$  is an  $n \times n$  matrix and  $\mathbf{b} \in \mathbb{R}^n$ , is called an *affine transformation*.

In general,  $F[\mathbf{x}]$  is an affine transformation if and only if  $L[\mathbf{x}] = F[\mathbf{x}] - F[\mathbf{0}]$  is a linear function. In the particular case (7.35),  $F[\mathbf{0}] = \mathbf{b}$ , and so  $L[\mathbf{x}] = A\mathbf{x}$ . The word “affine” comes from the Latin “affinus”, meaning “related”, because such transformations preserve the relation of parallelism between lines; see Exercise 7.3.2.

For example, every affine transformation from  $\mathbb{R}$  to  $\mathbb{R}$  has the form

$$f(x) = \alpha x + \beta. \quad (7.36)$$

As mentioned earlier, even though the graph of  $f(x)$  is a straight line,  $f$  is *not* a linear function — unless  $\beta = 0$ , and the line goes through the origin. Thus, to be mathematically accurate, we should refer to (7.36) as a *one-dimensional affine transformation*.

**Example 7.22.** The affine transformation

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

has the effect of first rotating the plane  $\mathbb{R}^2$  by  $90^\circ$  about the origin, and then translating by the vector  $(1, -2)^T$ . The reader may enjoy proving that this combination has the same effect as just rotating the plane through an angle of  $90^\circ$  centered at the point  $(\frac{3}{4}, -\frac{1}{2})$ . For details, see Exercise 7.3.14.

Note that the affine transformation (7.35) can be obtained by composing a linear function  $L[\mathbf{x}] = A\mathbf{x}$  with a translation  $T[\mathbf{x}] = \mathbf{x} + \mathbf{b}$ , so

$$F[\mathbf{x}] = T \circ L[\mathbf{x}] = T[L[\mathbf{x}]] = T[A\mathbf{x}] = A\mathbf{x} + \mathbf{b}.$$

The order of composition is important, since  $G = L \circ T$  defines the slightly different affine transformation

$$G[\mathbf{x}] = L \circ T[\mathbf{x}] = L[T[\mathbf{x}]] = L[\mathbf{x} + \mathbf{b}] = A(\mathbf{x} + \mathbf{b}) = A\mathbf{x} + \mathbf{c}, \quad \text{where } \mathbf{c} = A\mathbf{b}.$$

More generally, the composition of any two affine transformations is again an affine transformation. Specifically, given

$$F[\mathbf{x}] = A\mathbf{x} + \mathbf{a}, \quad G[\mathbf{y}] = B\mathbf{y} + \mathbf{b},$$

then

$$(G \circ F)[\mathbf{x}] = G[F[\mathbf{x}]] = G[A\mathbf{x} + \mathbf{a}] = B(A\mathbf{x} + \mathbf{a}) + \mathbf{b} = C\mathbf{x} + \mathbf{c}, \quad (7.37)$$

where  $C = BA$ ,  $\mathbf{c} = B\mathbf{a} + \mathbf{b}$ .

Note that the coefficient matrix of the composition is the product of the coefficient matrices, but the resulting vector of translation is *not* the sum of the two translation vectors.

## Exercises

7.3.1. *True or false:* An affine transformation takes (a) straight lines to straight lines; (b) triangles to triangles; (c) squares to squares; (d) circles to circles; (e) ellipses to ellipses.

◇ 7.3.2. (a) Let  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$  be an affine transformation. Let  $L_1, L_2 \subset \mathbb{R}^n$  be two parallel lines. Prove that  $F[L_1]$  and  $F[L_2]$  are also parallel lines.

(b) Is the converse valid: if  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$  maps parallel lines to parallel lines, then  $F$  is necessarily an affine transformation?

7.3.3. Describe the image of (i) the  $x$ -axis, (ii) the unit disk  $x^2 + y^2 \leq 1$ , (iii) the unit square  $0 \leq x, y \leq 1$ , under the following affine transformations:

$$(a) T_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad (b) T_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

$$(c) T_3 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad (d) T_4 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$(e) T_5 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} .6 & .8 \\ -.8 & .6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} -3 \\ 2 \end{pmatrix}, \quad (f) T_6 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$(g) T_7 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \quad (h) T_8 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

7.3.4. Using the affine transformations in Exercise 7.3.3, write out the following compositions and verify that they satisfy (7.37):

- (a)  $T_3 \circ T_4$ , (b)  $T_4 \circ T_3$ , (c)  $T_3 \circ T_6$ , (d)  $T_6 \circ T_3$ , (e)  $T_7 \circ T_8$ , (f)  $T_8 \circ T_7$ .

7.3.5. Describe the image of the triangle with vertices  $(-1, 0), (1, 0), (0, 2)$  under the affine

$$\text{transformation } T \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 & -1 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}.$$

7.3.6. Under what conditions is the composition of two affine transformations

- (a) a translation? (b) a linear function?

7.3.7. (a) Under what conditions does an affine transformation have an inverse? (b) Is the inverse an affine transformation? If so, find a formula for its matrix and vector constituents. (c) Find the inverse, when it exists, of each of the the affine transformations in Exercise 7.3.3.

◇ 7.3.8. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be a basis for  $\mathbb{R}^n$ . (a) Show that every affine transformation

$$F[\mathbf{x}] = A\mathbf{x} + \mathbf{b} \text{ on } \mathbb{R}^n \text{ is uniquely determined by the } n+1 \text{ vectors } \mathbf{w}_0 = F[\mathbf{0}], \mathbf{w}_1 = F[\mathbf{v}_1], \dots,$$

$\mathbf{w}_n = F[\mathbf{v}_n]$ . (b) Find the formula for  $A$  and  $\mathbf{b}$  when  $\mathbf{v}_1 = \mathbf{e}_1, \dots, \mathbf{v}_n = \mathbf{e}_n$  are the standard basis vectors. (c) Find the formula for  $A, \mathbf{b}$  for a general basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ .

7.3.9. Show that the space of all affine transformations on  $\mathbb{R}^n$  is a vector space. What is its dimension?

◇ 7.3.10. In this exercise, we establish a useful matrix representation for affine transformations. We identify  $\mathbb{R}^n$  with the  $n$ -dimensional affine subspace (as in Exercise 2.2.28)

$$V_n = \{ (\mathbf{x}, 1)^T = (x_1, \dots, x_n, 1)^T \} \subset \mathbb{R}^{n+1}$$

consisting of vectors whose last coordinate is fixed at  $x_{n+1} = 1$ . (a) Show that

$$\text{multiplication of vectors } \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \in V_n \text{ by the } (n+1) \times (n+1) \text{ affine matrix } \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{pmatrix}$$

coincides with the action (7.35) of an affine transformation on  $\mathbf{x} \in \mathbb{R}^n$ . (b) Prove that the composition law (7.37) for affine transformations corresponds to multiplication of their affine matrices. (c) Define the inverse of an affine transformation in the evident manner, and show that it corresponds to the inverse affine matrix.

## Isometry

A transformation that preserves distance is known as an *isometry*. (The mathematical term *metric* refers to the underlying norm or distance on the space; thus, “isometric” translates as “distance-preserving”.) In Euclidean geometry, the isometries coincide with the *rigid motions* — translations, rotations, reflections, and the affine maps they generate through composition.

**Definition 7.23.** Let  $V$  be a normed vector space. A function  $F: V \rightarrow V$  is called an *isometry* if it preserves distance, meaning

$$d(F[\mathbf{v}], F[\mathbf{w}]) = d(\mathbf{v}, \mathbf{w}) \quad \text{for all } \mathbf{v}, \mathbf{w} \in V. \quad (7.38)$$

Since the distance between points is just the norm of the vector connecting them,  $d(\mathbf{v}, \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|$ , cf. (3.33), the isometry condition (7.38) can be restated as

$$\|F[\mathbf{v}] - F[\mathbf{w}]\| = \|\mathbf{v} - \mathbf{w}\| \quad \text{for all } \mathbf{v}, \mathbf{w} \in V. \quad (7.39)$$

Clearly, any translation

$$T[\mathbf{v}] = \mathbf{v} + \mathbf{a}, \quad \text{where } \mathbf{a} \in V,$$

defines an isometry, since  $T[\mathbf{v}] - T[\mathbf{w}] = \mathbf{v} - \mathbf{w}$ . A linear transformation  $L: V \rightarrow V$  defines an isometry if and only if

$$\|L[\mathbf{v}]\| = \|\mathbf{v}\| \quad \text{for all } \mathbf{v} \in V, \quad (7.40)$$

because, by linearity,

$$\|L[\mathbf{v}] - L[\mathbf{w}]\| = \|L[\mathbf{v} - \mathbf{w}]\| = \|\mathbf{v} - \mathbf{w}\|.$$

A similar computation proves that an affine transformation  $F[\mathbf{v}] = L[\mathbf{v}] + \mathbf{a}$  is an isometry if and only if its linear part  $L[\mathbf{v}]$  is.

As noted above, the simplest class of isometries comprises the translations

$$T[\mathbf{x}] = \mathbf{x} + \mathbf{b} \quad (7.41)$$

in the direction  $\mathbf{b}$ . For the standard Euclidean norm on  $V = \mathbb{R}^n$ , the linear isometries consist of rotations and reflections. As we shall prove, both are characterized by orthogonal matrices:

$$L[\mathbf{x}] = Q\mathbf{x}, \quad \text{where} \quad Q^T Q = \mathbf{I}. \quad (7.42)$$

The *proper isometries* correspond to the rotations, with  $\det Q = +1$ , and can be realized as physical motions; *improper isometries*, with  $\det Q = -1$ , are then obtained by reflection in a mirror.

**Proposition 7.24.** A linear transformation  $L[\mathbf{x}] = Q\mathbf{x}$  defines a Euclidean isometry of  $\mathbb{R}^n$  if and only if  $Q$  is an orthogonal matrix.

*Proof:* The linear isometry condition (7.40) requires that

$$\|Q\mathbf{x}\|^2 = (Q\mathbf{x})^T Q\mathbf{x} = \mathbf{x}^T Q^T Q\mathbf{x} = \mathbf{x}^T \mathbf{x} = \|\mathbf{x}\|^2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

According to Exercise 4.3.16, this holds if and only if  $Q^T Q = \mathbf{I}$ , which is precisely the condition (4.29) that  $Q$  be an orthogonal matrix. *Q.E.D.*

It can be proved, [93], that the most general Euclidean isometry of  $\mathbb{R}^n$  is an affine transformation, and hence of the form  $F[\mathbf{x}] = Q\mathbf{x} + \mathbf{b}$ , where  $Q$  is an orthogonal matrix and  $\mathbf{b}$  is a vector. Therefore, every Euclidean isometry or rigid motion is a combination of translations, rotations, and reflections.

In the two-dimensional case, the proper linear isometries  $R[\mathbf{x}] = Q\mathbf{x}$  with  $\det Q = 1$  represent rotations around the origin. More generally, a rotation of the plane around a center at  $\mathbf{c}$  is represented by the affine isometry

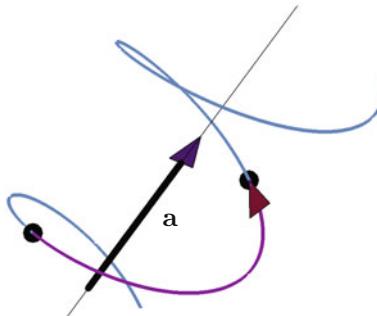
$$R[\mathbf{x}] = Q(\mathbf{x} - \mathbf{c}) + \mathbf{c} = Q\mathbf{x} + \mathbf{b}, \quad \text{where} \quad \mathbf{b} = (\mathbf{I} - Q)\mathbf{c}, \quad (7.43)$$

and where  $Q$  is a rotation matrix. In Exercise 7.3.14, we ask you to prove that every plane isometry is either a translation or a rotation around a center.

In three-dimensional space, both translations (7.41) and rotations around a center (7.43) continue to define proper isometries. There is one additional type, representing the motion of a point on the head of a screw. A *screw motion* is an affine transformation of the form

$$S[\mathbf{x}] = Q\mathbf{x} + \mathbf{a}, \quad (7.44)$$

where the  $3 \times 3$  orthogonal matrix  $Q$  represents a rotation through an angle  $\theta$  around a fixed axis in the direction of the vector  $\mathbf{a}$ , which is also the direction of the translation



**Figure 7.11.** A Screw Motion.

term. The result is indicated in [Figure 7.11](#); the trajectory followed by a point not on the axis is a circular helix centered on the axis. For example,

$$S_\theta \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}$$

represents a vertical screw along the  $z$ -axis through an angle  $\theta$  by a distance  $a$ . In Exercise 8.2.45 you are asked to prove that every proper isometry of  $\mathbb{R}^3$  is either a translation, a rotation, or a screw motion.

The isometries of  $\mathbb{R}^2$  and  $\mathbb{R}^3$  are indispensable for understanding of how physical objects move in three-dimensional space. Basic computer graphics and animation require efficient implementation of rigid isometries in three-dimensional space and their compositions — coupled with appropriate (nonlinear) perspective maps prescribing the projection of three-dimensional objects onto a two-dimensional viewing screen, [\[12, 72\]](#).

## Exercises

*Note:* All exercises are based on the Euclidean norm unless otherwise noted.

7.3.11. Which of the indicated maps  $\mathbf{F}(x, y)$  define isometries of the Euclidean plane?

(a)  $\begin{pmatrix} y \\ -x \end{pmatrix}$ , (b)  $\begin{pmatrix} x-2 \\ y-1 \end{pmatrix}$ , (c)  $\begin{pmatrix} x-y+1 \\ x+2 \end{pmatrix}$ , (d)  $\frac{1}{\sqrt{2}} \begin{pmatrix} x+y-3 \\ x+y-2 \end{pmatrix}$ , (e)  $\frac{1}{5} \begin{pmatrix} 3x+4y \\ -4x+3y+1 \end{pmatrix}$ .

7.3.12. Prove that the planar affine isometry  $F \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y+1 \\ x-2 \end{pmatrix}$  represents a rotation through an angle of  $90^\circ$  around the center  $\left(\frac{3}{2}, -\frac{1}{2}\right)^T$ .

7.3.13. *True or false:* The map  $L[\mathbf{x}] = -\mathbf{x}$  for  $\mathbf{x} \in \mathbb{R}^n$  defines (a) an isometry; (b) a rotation.

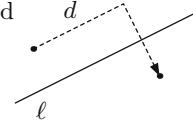
◇ 7.3.14. Prove that every *proper* affine plane isometry  $F[\mathbf{x}] = Q\mathbf{x} + \mathbf{b}$  of  $\mathbb{R}^2$ , where  $\det Q = 1$ , is either (i) a translation, or (ii) a rotation (7.43) centered at some point  $\mathbf{c} \in \mathbb{R}^2$ .

*Hint:* Use Exercise 1.5.7.

7.3.15. Compute both compositions  $F \circ G$  and  $G \circ F$  of the following affine transformations on  $\mathbb{R}^2$ . Which pairs commute? (a)  $F$  = counterclockwise rotation around the origin by  $45^\circ$ ;  $G$  = translation in the  $y$  direction by 3 units. (b)  $F$  = counterclockwise rotation around the point  $(1, 1)^T$  by  $30^\circ$ ;  $G$  = counterclockwise rotation around the point  $(-2, 1)^T$  by  $90^\circ$ . (c)  $F$  = reflection through the line  $y = x + 1$ ;  $G$  = rotation around  $(1, 1)^T$  by  $180^\circ$ .

◇ 7.3.16. In  $\mathbb{R}^2$ , show the following: (a) The composition of two affine isometries is another affine isometry. (b) The composition of two translations is another translation. (c) The composition of a translation and a rotation (not necessarily centered at the origin) in either order is a rotation. (d) The composition of two plane rotations is either another rotation or a translation. What is the condition for the latter possibility? (e) Every plane translation can be written as the composition of two rotations.

◇ 7.3.17. Let  $\ell$  be a line in  $\mathbb{R}^2$ . A *glide reflection* is an affine map on  $\mathbb{R}^2$  composed of a translation in the direction of  $\ell$  by a distance  $d$  followed by a reflection through  $\ell$ . Find the formula for a glide reflection along (a) the  $x$ -axis by a distance 2; (b) the line  $y = x$  by a distance 3 in the direction of increasing  $x$ ; (c) the line  $x + y = 1$  by a distance 2 in the direction of increasing  $x$ .



◇ 7.3.18. Let  $\ell$  be the line in the direction of the unit vector  $\mathbf{u}$  through the point  $\mathbf{a}$ . (a) Write down the formula for the affine map defining the reflection through the line  $\ell$ . Hint: Use Exercise 7.2.12. (b) Write down the formula for the glide reflection, as defined in Exercise 7.3.17, along  $\ell$  by a distance  $d$  in the direction of  $\mathbf{u}$ . (c) Prove that every improper affine plane isometry is either a reflection or a glide reflection. Hint: Use Exercise 7.2.10.

◇ 7.3.19. A set of  $n + 1$  points  $\mathbf{a}_0, \dots, \mathbf{a}_n \in \mathbb{R}^n$  is said to be *in general position* if the differences  $\mathbf{a}_i - \mathbf{a}_j$  span  $\mathbb{R}^n$ . (a) Show that the points are in general position if and only if they do not all lie in a proper affine subspace  $A \subsetneq \mathbb{R}^n$ , cf. Exercise 2.2.28. (b) Let  $\mathbf{a}_0, \dots, \mathbf{a}_n$  and  $\mathbf{b}_0, \dots, \mathbf{b}_n$  be two sets in general position. Show that there is an isometry  $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$  such that  $F[\mathbf{a}_i] = \mathbf{b}_i$  for all  $i = 0, \dots, n$ , if and only if their interpoint distances agree:  $\|\mathbf{a}_i - \mathbf{a}_j\| = \|\mathbf{b}_i - \mathbf{b}_j\|$  for all  $0 \leq i < j \leq n$ . Hint: Use Exercise 4.3.19.

◇ 7.3.20. Suppose that  $V$  is an inner product space and  $L: V \rightarrow V$  is an isometry, so  $\|L[\mathbf{v}]\| = \|\mathbf{v}\|$  for all  $\mathbf{v} \in V$ . Prove that  $L$  also preserves the inner product:  $\langle L[\mathbf{v}], L[\mathbf{w}] \rangle = \langle \mathbf{v}, \mathbf{w} \rangle$ . Hint: Look at  $\|L[\mathbf{v} + \mathbf{w}]\|^2$ .

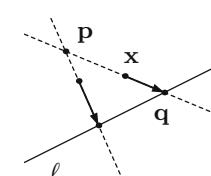
◇ 7.3.21. Let  $V$  be a normed vector space. Prove that a linear map  $L: V \rightarrow V$  defines an isometry of  $V$  for the given norm if and only if it maps the unit sphere  $S_1 = \{\|\mathbf{u}\| = 1\}$  to itself:  $L[S_1] = \{L[\mathbf{u}] \mid \mathbf{u} \in S_1\} = S_1$ .

7.3.22. (a) List all linear and affine isometries of  $\mathbb{R}^2$  with respect to the  $\infty$  norm. Hint: Use Exercise 7.3.21. (b) Can you generalize your results to  $\mathbb{R}^3$ ?

7.3.23. Answer Exercise 7.3.22 for the 1 norm.

◇ 7.3.24. A matrix of the form  $H = \begin{pmatrix} \cosh \alpha & \sinh \alpha \\ \sinh \alpha & \cosh \alpha \end{pmatrix}$  for  $\alpha \in \mathbb{R}$  defines a *hyperbolic rotation* of  $\mathbb{R}^2$ . (a) Prove that all hyperbolic rotations preserve the indefinite quadratic form  $q(\mathbf{x}) = x^2 - y^2$  in the sense that  $q(H\mathbf{x}) = q(\mathbf{x})$  for all  $\mathbf{x} = (x, y)^T \in \mathbb{R}^2$ . Observe that ordinary rotations preserve circles  $x^2 + y^2 = a$ , while hyperbolic rotations preserve hyperbolas  $x^2 - y^2 = a$ . (b) Are there any other affine transformations of  $\mathbb{R}^2$  that preserve the quadratic form  $q(\mathbf{x})$ ? **Remark.** The four-dimensional version of this construction, i.e., affine maps preserving the indefinite *Minkowski form*  $t^2 - x^2 - y^2 - z^2$ , forms the geometrical foundation for Einstein's theory of special relativity, [55].

◇ 7.3.25. Let  $\ell \subset \mathbb{R}^2$  be a line, and  $\mathbf{p} \notin \ell$  a point. A *perspective map* takes a point  $\mathbf{x} \in \mathbb{R}^2$  to the point  $\mathbf{q} \in \ell$  that is the intersection of  $\ell$  with the line going through  $\mathbf{p}$  and  $\mathbf{x}$ . If the line is parallel to  $\ell$ , then the map is not defined. Find the formula for the perspective map when (a)  $\ell$  is the  $x$ -axis and  $\mathbf{p} = (0, 1)^T$ , (b)  $\ell$  is the line  $y = x$  and  $\mathbf{p} = (1, 0)^T$ . Is either map affine? An isometry? **Remark.** Mapping three-dimensional objects onto a two-dimensional screen (or your retina) is based on perspective maps, which are thus of fundamental importance in art, optics, computer vision, computer graphics and animation, and computer games.



## 7.4 Linear Systems

The abstract notion of a linear system serves to unify, in a common conceptual framework, linear systems of algebraic equations, linear differential equations, both ordinary and partial, linear boundary value problems, linear integral equations, linear control systems, and a huge variety of other linear systems that appear in all aspects of mathematics and its applications. The idea is simply to replace matrix multiplication by a general linear function. Many of the structural results we learned in the matrix context have, when suitably formulated, direct counterparts in these more general frameworks. The result is a unified understanding of the basic properties and nature of solutions to all such linear systems.

**Definition 7.25.** A *linear system* is an equation of the form

$$L[\mathbf{u}] = \mathbf{f}, \quad (7.45)$$

in which  $L:U \rightarrow V$  is a linear function between vector spaces, the right-hand side is an element of the codomain,  $\mathbf{f} \in V$ , while the desired solution belongs to the domain,  $\mathbf{u} \in U$ . The system is *homogeneous* if  $\mathbf{f} = \mathbf{0}$ ; otherwise, it is called *inhomogeneous*.

**Example 7.26.** If  $U = \mathbb{R}^n$  and  $V = \mathbb{R}^m$ , then, according to Theorem 7.5, every linear function  $L:\mathbb{R}^n \rightarrow \mathbb{R}^m$  is given by matrix multiplication:  $L[\mathbf{u}] = A\mathbf{u}$ . Therefore, in this particular case, every linear system is a matrix system, namely  $A\mathbf{u} = \mathbf{f}$ .

**Example 7.27.** A *linear ordinary differential equation* takes the form  $L[u] = f$ , where  $L$  is an  $n^{\text{th}}$  order linear differential operator of the form (7.15), and the right-hand side is, say, a continuous function. Written out, the differential equation takes the familiar form

$$L[u] = a_n(x) \frac{d^n u}{dx^n} + a_{n-1}(x) \frac{d^{n-1} u}{dx^{n-1}} + \cdots + a_1(x) \frac{du}{dx} + a_0(x)u = f(x). \quad (7.46)$$

You should have already gained some familiarity with solving the constant coefficient case as covered, for instance, in [7, 22].

**Example 7.28.** Let  $K(x, y)$  be a function of two variables that is continuous for all  $a \leq x, y \leq b$ . Then the integral

$$I_K[u] = \int_a^b K(x, y) u(y) dy$$

defines a linear operator  $I_K: C^0[a, b] \rightarrow C^0[a, b]$ , known as an *integral transform*. Important examples include the Fourier and Laplace transforms, [61, 79]. Finding the inverse transform requires solving a *linear integral equation*  $I_K[u] = f$ , which has the explicit form

$$\int_a^b K(x, y) u(y) dy = f(x).$$

**Example 7.29.** We can combine linear maps to form more complicated, “mixed” types of linear systems. For example, consider a typical initial value problem

$$u'' + u' - 2u = x, \quad u(0) = 1, \quad u'(0) = -1, \quad (7.47)$$

for an unknown scalar function  $u(x)$ . The differential equation can be written as a linear system

$$L[u] = x, \quad \text{where} \quad L[u] = (D^2 + D - 2)[u] = u'' + u' - 2u$$

is a linear, constant coefficient differential operator. Further,

$$M[u] = \begin{pmatrix} L[u] \\ u(0) \\ u'(0) \end{pmatrix} = \begin{pmatrix} u''(x) + u'(x) - 2u(x) \\ u(0) \\ u'(0) \end{pmatrix}$$

defines a linear map whose domain is the space  $U = C^2$  of twice continuously differentiable functions  $u(x)$ , and whose image is the vector space  $V$  consisting of all triples<sup>†</sup>  $\mathbf{v} = \begin{pmatrix} f(x) \\ a \\ b \end{pmatrix}$ , where  $f \in C^0$  is a continuous function and  $a, b \in \mathbb{R}$  are real constants. You

should convince yourself that  $V$  is indeed a vector space under the evident addition and scalar multiplication operations. In this way, we can write the initial value problem (7.47) in linear systems form as  $M[u] = \mathbf{f}$ , where  $\mathbf{f} = (x, 1, -1)^T$ .

A similar construction applies to linear boundary value problems. For example, the boundary value problem

$$u'' + u = e^x, \quad u(0) = 1, \quad u(1) = 2,$$

is in the form of a linear system

$$B[u] = \mathbf{f}, \quad \text{where} \quad B[u] = \begin{pmatrix} u''(x) + u(x) \\ u(0) \\ u(1) \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} e^x \\ 1 \\ 2 \end{pmatrix}.$$

Note that  $B: C^2 \rightarrow V$  defines a linear map having the same domain and codomain as the initial value problem map  $M$ .

## Exercises

7.4.1. *True or false:* If  $F[\mathbf{x}]$  is an affine transformation on  $\mathbb{R}^n$ , then the equation  $F[\mathbf{x}] = \mathbf{c}$  defines a linear system.

7.4.2. Place each of the following linear systems in the form (7.45). Carefully describe the linear function, its domain, its codomain, and the right-hand side of the system. Which systems are homogeneous? (a)  $3x + 5 = 0$ , (b)  $x = y + z$ , (c)  $a = 2b - 3$ ,  $b = c - 1$ , (d)  $3(p-2) = 2(q-3)$ ,  $p+q = 0$ , (e)  $u' + 3xu = 0$ , (f)  $u' + 3x = 0$ , (g)  $u' = u$ ,  $u(0) = 1$ , (h)  $u'' - u = e^x$ ,  $u(0) = 3u(1)$ , (i)  $u'' + x^2u = 3x$ ,  $u(0) = 1$ ,  $u'(0) = 0$ , (j)  $u' = v$ ,  $v' = 2u$ , (k)  $u'' - v'' = 2u - v$ ,  $u(0) = v(0)$ ,  $u(1) = v(1)$ , (l)  $u(x) = 1 - 3 \int_0^x u(y) dy$ , (m)  $\int_0^\infty u(t) e^{-st} dt = 1 + s^2$ , (n)  $\int_0^1 u(x) dx = u\left(\frac{1}{2}\right)$ , (o)  $\int_0^1 u(y) dy = \int_0^1 y v(y) dy$ , (p)  $\frac{\partial u}{\partial t} + 2 \frac{\partial u}{\partial x} = 1$ , (q)  $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ ,  $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$ , (r)  $-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = x^2 + y^2 - 1$ .

7.4.3. The Fredholm Alternative of Theorem 4.46 first appeared in the study of what are now known as *Fredholm integral equations*:  $u(x) + \int_a^b K(x, y) u(y) dy = f(x)$ , in which  $K(x, y)$  and  $f(x)$  are prescribed continuous functions. Explain how the integral equation is a linear system; i.e., describe the linear map  $L$ , its domain and codomain, and prove linearity.

<sup>†</sup> This is a particular case of the general Cartesian product construction between vector spaces; here  $V = C^0 \times \mathbb{R}^2$ . See Exercise 2.1.13 for details.

7.4.4. Answer Exercise 7.4.3 for the *Volterra integral equation*  $u(t) + \int_a^t K(t, s) u(s) ds = f(t)$ , where  $a \leq t \leq b$ .

7.4.5. (a) Prove that the solution to the linear integral equation  $u(t) = a + \int_0^t k(s) u(s) ds$  solves the linear initial value problem  $\frac{du}{dt} = k(t) u(t)$ ,  $u(0) = a$ .

(b) Use part (a) to solve the following integral equations

$$(i) \ u(t) = 2 - \int_0^t u(s) ds, \quad (ii) \ u(t) = 1 + 2 \int_1^t s u(s) ds, \quad (iii) \ u(t) = 3 + \int_0^t e^s u(s) ds.$$

## The Superposition Principle

Before attempting to tackle general inhomogeneous linear systems, we should look first at the homogeneous version. The most important fact is that homogeneous linear systems admit a superposition principle that allows one to construct new solutions from known solutions. Recall that the word “superposition” refers to taking linear combinations of solutions.

Consider a general homogeneous linear system

$$L[\mathbf{z}] = \mathbf{0}, \quad (7.48)$$

where  $L: U \rightarrow V$  is a linear function. If we are given two solutions, say  $\mathbf{z}_1$  and  $\mathbf{z}_2$ , meaning that

$$L[\mathbf{z}_1] = \mathbf{0}, \quad L[\mathbf{z}_2] = \mathbf{0},$$

then their sum  $\mathbf{z}_1 + \mathbf{z}_2$  is automatically a solution, since, in view of the linearity of  $L$ ,

$$L[\mathbf{z}_1 + \mathbf{z}_2] = L[\mathbf{z}_1] + L[\mathbf{z}_2] = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

Similarly, given a solution  $\mathbf{z}$  and any scalar  $c$ , the scalar multiple  $c\mathbf{z}$  is automatically a solution, since

$$L[c\mathbf{z}] = cL[\mathbf{z}] = c\mathbf{0} = \mathbf{0}.$$

Combining these two elementary observations, we can now state the general *superposition principle*. The proof is an immediate consequence of formula (7.4).

**Theorem 7.30.** If  $\mathbf{z}_1, \dots, \mathbf{z}_k$  are all solutions to the same homogeneous linear system  $L[\mathbf{z}] = \mathbf{0}$ , then every linear combination  $c_1\mathbf{z}_1 + \dots + c_k\mathbf{z}_k$  is also a solution.

As with matrices, we call the solution space to the homogeneous linear system (7.48) the *kernel* of the linear function  $L$ . The superposition principle implies that the kernel always forms a subspace.

**Proposition 7.31.** If  $L: U \rightarrow V$  is a linear function, then its *kernel*

$$\ker L = \{ \mathbf{z} \in U \mid L[\mathbf{z}] = \mathbf{0} \} \subset U \quad (7.49)$$

is a subspace of the domain space  $U$ .

As we know, in the case of linear matrix systems, the kernel can be explicitly determined by applying the usual Gaussian Elimination algorithm. To solve more general homogeneous linear systems, e.g., linear differential equations, one must develop appropriate analytical solution techniques.

**Example 7.32.** Consider the second order linear differential operator

$$L = D^2 - 2D - 3, \quad (7.50)$$

which maps the function  $u(x)$  to the function

$$L[u] = (D^2 - 2D - 3)[u] = u'' - 2u' - 3u.$$

The associated homogeneous system takes the form of a homogeneous, linear, constant coefficient second order ordinary differential equation

$$L[u] = u'' - 2u' - 3u = 0. \quad (7.51)$$

In accordance with the standard solution method, we plug the exponential ansatz<sup>†</sup>

$$u = e^{\lambda x}$$

into the equation. The result is

$$L[e^{\lambda x}] = D^2[e^{\lambda x}] - 2D[e^{\lambda x}] - 3e^{\lambda x} = \lambda^2 e^{\lambda x} - 2\lambda e^{\lambda x} - 3e^{\lambda x} = (\lambda^2 - 2\lambda - 3)e^{\lambda x}.$$

Therefore,  $u = e^{\lambda x}$  is a solution if and only if  $\lambda$  satisfies the *characteristic equation*

$$0 = \lambda^2 - 2\lambda - 3 = (\lambda - 3)(\lambda + 1).$$

The two roots are  $\lambda_1 = 3$ ,  $\lambda_2 = -1$ , and hence

$$u_1(x) = e^{3x}, \quad u_2(x) = e^{-x}, \quad (7.52)$$

are two linearly independent solutions of (7.51). According to the general superposition principle, every linear combination

$$u(x) = c_1 u_1(x) + c_2 u_2(x) = c_1 e^{3x} + c_2 e^{-x} \quad (7.53)$$

of these two basic solutions is also a solution, for any choice of constants  $c_1, c_2$ . In fact, this two-parameter family (7.53) constitutes the most general solution to the ordinary differential equation (7.51); indeed, this is a consequence of Theorem 7.34 below. Thus, the kernel of the second order differential operator (7.50) is two-dimensional, with basis given by the independent exponential solutions (7.52).

In general, the solution space to an  $n^{\text{th}}$  order homogeneous linear ordinary differential equation

$$L[u] = a_n(x) \frac{d^n u}{dx^n} + a_{n-1}(x) \frac{d^{n-1} u}{dx^{n-1}} + \cdots + a_1(x) \frac{du}{dx} + a_0(x)u = 0 \quad (7.54)$$

is a subspace of the vector space  $C^n(a, b)$  of  $n$  times continuously differentiable functions defined on an open interval<sup>‡</sup>  $a < x < b$ , since it is just the kernel of a linear differential

<sup>†</sup> The German word *Ansatz* refers to the method of finding a solution to a complicated equation by guessing the solution's form in advance. Typically, one is not clever enough to guess the precise solution, and so the ansatz will have one or more free parameters — in this case the constant exponent  $\lambda$  — that, with some luck, can be rigged up to fulfill the requirements imposed by the equation. Thus, a reasonable English translation of “ansatz” is “inspired guess”.

<sup>‡</sup> We allow  $a$  and/or  $b$  to be infinite.

operator  $L: C^n(a, b) \rightarrow C^0(a, b)$ . This implies that linear combinations of solutions are also solutions. To determine the number of solutions, or, more precisely, the dimension of the solution space, we need to impose some mild restrictions on the differential operator.

**Definition 7.33.** A differential operator  $L$  given by (7.54) is called *nonsingular* on an open interval  $(a, b)$  if all its coefficients are continuous functions, so  $a_n(x), \dots, a_0(x) \in C^0(a, b)$ , and its *leading coefficient* does not vanish:  $a_n(x) \neq 0$  for all  $a < x < b$ .

The basic existence and uniqueness theorems governing nonsingular homogeneous linear ordinary differential equations can be reformulated as a characterization of the dimension of the solution space.

**Theorem 7.34.** The kernel of a nonsingular  $n^{\text{th}}$  order ordinary differential operator is an  $n$ -dimensional subspace  $\ker L \subset C^n(a, b)$ .

A proof of this theorem relies on the fundamental existence and uniqueness theorems for ordinary differential equations, and can be found in [7, 36]. The fact that the kernel has dimension  $n$  means that it has a basis consisting of  $n$  linearly independent solutions  $u_1(x), \dots, u_n(x) \in C^n(a, b)$  with the property that every solution to the homogeneous differential equation (7.54) is given by a linear combination

$$u(x) = c_1 u_1(x) + \dots + c_n u_n(x),$$

where  $c_1, \dots, c_n$  are arbitrary constants. Therefore, once we find  $n$  linearly independent solutions of an  $n^{\text{th}}$  order homogeneous linear ordinary differential equation, we can immediately write down its most general solution.

The condition that the leading coefficient  $a_n(x) \neq 0$  is essential. Points where  $a_n(x) = 0$  are known as *singular points*. Singular points show up in many applications, and must be treated separately and with care, [7, 22, 61]. Of course, if the coefficients are constant, then there is nothing to worry about — either the leading coefficient is nonzero,  $a_n \neq 0$ , or the differential equation is, in fact, of lower order than advertised. Here is the prototypical example of an ordinary differential equation with a singular point.

**Example 7.35.** A second order *Euler differential equation* takes the form

$$E[u] = ax^2 u'' + bxu' + cu = 0, \quad (7.55)$$

where  $a \neq 0$  and  $b, c$  are constants. Here  $E = ax^2 D^2 + bx D + c$  is a second order variable coefficient linear differential operator. Instead of the exponential solution ansatz used in the constant coefficient case, Euler equations are solved by using a power ansatz

$$u(x) = x^r \quad (7.56)$$

with unknown exponent  $r$ . Substituting into the differential equation, we find

$$\begin{aligned} E[x^r] &= ax^2 D^2[x^r] + bx D[x^r] + cx^r \\ &= ar(r-1)x^r + brx^r + cx^r = [ar(r-1) + br + c]x^r. \end{aligned}$$

Thus,  $x^r$  is a solution if and only if  $r$  satisfies the *characteristic equation*

$$ar(r-1) + br + c = ar^2 + (b-a)r + c = 0. \quad (7.57)$$

If the quadratic characteristic equation has two distinct real roots,  $r_1 \neq r_2$ , then we obtain two linearly independent solutions  $u_1(x) = x^{r_1}$  and  $u_2(x) = x^{r_2}$ , and so the general (real)

solution to (7.55) has the form

$$u(x) = c_1 |x|^{r_1} + c_2 |x|^{r_2}. \quad (7.58)$$

(The absolute values are usually needed to ensure that the solutions remain real when  $x < 0$ .) The other cases — repeated roots and complex roots — will be discussed below.

The Euler equation has a singular point at  $x = 0$ , where its leading coefficient vanishes. Theorem 7.34 assures us that the differential equation has a two-dimensional solution space on every interval not containing the singular point. However, predicting the number of solutions that remain continuously differentiable at  $x = 0$  is not so easy, since it depends on the values of the exponents  $r_1$  and  $r_2$ . For instance, the case

$$x^2 u'' - 3x u' + 3u = 0 \quad \text{has general solution} \quad u = c_1 x + c_2 x^3,$$

which forms a two-dimensional subspace of  $C^0(\mathbb{R})$ . However,

$$x^2 u'' + x u' - u = 0 \quad \text{has general solution} \quad u = c_1 x + \frac{c_2}{x},$$

and only the multiples of the first solution  $x$  are continuous at  $x = 0$ . Therefore, the solutions that are continuous everywhere form only a one-dimensional subspace of  $C^0(\mathbb{R})$ . Finally,

$$x^2 u'' + 5x u' + 3u = 0 \quad \text{has general solution} \quad u = \frac{c_1}{x} + \frac{c_2}{x^3}.$$

In this case, there are no nontrivial solutions  $u(x) \not\equiv 0$  that are continuous at  $x = 0$ , and so the space of solutions defined on all of  $\mathbb{R}$  is zero-dimensional.

The superposition principle is equally valid in the study of homogeneous linear partial differential equations. Here is a particularly noteworthy example.

**Example 7.36.** Consider the *Laplace equation*

$$\Delta[u] = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (7.59)$$

for a function  $u(x, y)$  defined on a domain  $\Omega \subset \mathbb{R}^2$ . The Laplace equation is named after the renowned eighteenth-century French mathematician Pierre-Simon Laplace, and is the most important partial differential equation. Its applications range over almost all fields of mathematics, physics, and engineering, including complex analysis, differential geometry, fluid mechanics, electromagnetism, elasticity, thermodynamics, and quantum mechanics, [61]. The Laplace equation is a homogeneous linear partial differential equation corresponding to the partial differential operator  $\Delta = \partial_x^2 + \partial_y^2$  known as the *Laplacian*. Linearity can either be proved directly, or by noting that  $\Delta$  is built up from the basic linear partial derivative operators  $\partial_x, \partial_y$  by the processes of composition and addition, as detailed in Exercise 7.1.46. Solutions to the Laplace equation are known as *harmonic functions*.

Unlike homogeneous linear ordinary differential equations, there is an infinite number of linearly independent solutions to the Laplace equation. Examples include the trigonometric/exponential solutions

$$e^{\omega x} \cos \omega y, \quad e^{\omega x} \sin \omega y, \quad e^{\omega y} \cos \omega x, \quad e^{\omega y} \sin \omega y,$$

where  $\omega$  is any real constant. There are also infinitely many independent *harmonic polynomial* solutions, the first few of which are

$$1, \quad x, \quad y, \quad x^2 - y^2, \quad xy, \quad x^3 - 3xy^2, \quad \dots.$$

The reader might enjoy finding some more polynomial solutions and trying to spot the pattern. (The answer will appear shortly.) As usual, we can build up more complicated solutions by taking general linear combinations of these particular ones; for instance,  $u(x, y) = 1 - 4xy + 2e^{3x} \cos 3y$  is automatically a solution. See [61] for further developments.

## Exercises

- 7.4.6. Solve the following homogeneous linear ordinary differential equations. What is the dimension of the solution space? (a)  $u'' - 4u = 0$ , (b)  $u'' - 6u' + 8u = 0$ ,  
 (c)  $u''' - 9u' = 0$ , (d)  $u'''' + 4u''' - u'' - 16u' - 12u = 0$ .
- 7.4.7. Define  $L[y] = y'' + y$ . (a) Prove directly from the definition that  $L: C^2[a, b] \rightarrow C^0[a, b]$  is a linear transformation. (b) Determine  $\ker L$ .
- 7.4.8. Answer Exercise 7.4.7 when  $L = 3D^2 - 2D - 5$ .
- 7.4.9. Consider the linear differential equation  $y''' + 5y'' + 3y' - 9y = 0$ . (a) Write the equation in the form  $L[y] = 0$  for a differential operator  $L = p(D)$ . (b) Find a basis for  $\ker L$ , and then write out the general solution to the differential equation.
- 7.4.10. The following functions are solutions to a constant coefficient homogeneous scalar ordinary differential equation. (i) Determine the least possible order of the differential equation, and (ii) write down an appropriate differential equation.  
 (a)  $e^{2x} + e^{-3x}$ , (b)  $1 + e^{-x}$ , (c)  $xe^x$ , (d)  $e^x + 2e^{2x} + 3e^{3x}$ .
- 7.4.11. Solve the following Euler differential equations: (a)  $x^2u'' + 5xu' - 5u = 0$ ,  
 (b)  $2x^2u'' - xu' - 2u = 0$ , (c)  $x^2u'' - u = 0$ , (d)  $x^2u'' + xu' - 3u = 0$ ,  
 (e)  $3x^2u'' - 5xu' - 3u = 0$ , (f)  $\frac{d^2u}{dx^2} + \frac{2}{x}\frac{du}{dx} = 0$ .
- 7.4.12. Solve the third order Euler differential equation  $x^3u''' + 2x^2u'' - 3xu' + 3u = 0$  by using the power ansatz (7.56). What is the dimension of the solution space for  $x > 0$ ? For all  $x$ ?
- 7.4.13. (i) Show that if  $u(x)$  solves the Euler equation  $ax^2\frac{d^2u}{dx^2} + bx\frac{du}{dx} + cu = 0$ , then  $v(t) = u(e^t)$  solves a linear, constant coefficient differential equation. (ii) Use this alternative technique to solve the Euler differential equations in Exercise 7.4.11.
- ◇ 7.4.14. (a) Use the method in Exercise 7.4.13 to solve an Euler equation whose characteristic equation has a double root  $r_1 = r_2 = r$ . (b) Solve the specific equations  
 (i)  $x^2u'' - xu' + u = 0$ , (ii)  $\frac{d^2u}{dx^2} + \frac{1}{x}\frac{du}{dx} = 0$ .
- 7.4.15. Show that if  $u(x)$  solves  $xu'' + 2u' - 4xu = 0$ , then  $v(x) = xu(x)$  solves a linear, constant coefficient equation. Use this to find the general solution to the given differential equation. Which of your solutions are continuous at the singular point  $x = 0$ ? Differentiable?
- 7.4.16. Let  $S \subset \mathbb{R}$  be an open subset (i.e., a union of open intervals), and let  $D: C^1(S) \rightarrow C^0(S)$  be the derivative operator  $D[f] = f'$ . True or false:  $\ker D$  is a one-dimensional subspace of  $C^1(S)$ .
- 7.4.17. Show that  $\log(x^2 + y^2)$  and  $\frac{x}{x^2 + y^2}$  are harmonic functions, that is, solutions of the two-dimensional Laplace equation.

- 7.4.18. Find all solutions  $u = f(r)$  of the two-dimensional Laplace equation that depend only on the radial coordinate  $r = \sqrt{x^2 + y^2}$ . Do these solutions form a vector space? If so, what is its dimension?
- 7.4.19. Find all (real) solutions to the two-dimensional Laplace equation of the form  $u = \log p(x, y)$ , where  $p(x, y)$  is a quadratic polynomial. Do these solutions form a vector space? If so, what is its dimension?
- 7.4.20. (a) Show that the function  $e^x \cos y$  is a solution to the two-dimensional Laplace equation. (b) Show that its quadratic Taylor polynomial at  $x = y = 0$  is harmonic. (c) What about its degree 3 Taylor polynomial? (d) Can you state a general theorem? (e) Test your result by looking at the Taylor polynomials of the harmonic function  $\log[(x-1)^2 + y^2]$ .
- 7.4.21. (a) Find a basis for, and the dimension of, the vector space consisting of all quadratic polynomial solutions of the three-dimensional Laplace equation  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$ . (b) Do the same for the homogeneous cubic polynomial solutions.
- 7.4.22. Find all solutions  $u = f(r)$  of the three-dimensional Laplace equation  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$  that depend only on the radial coordinate  $r = \sqrt{x^2 + y^2 + z^2}$ . Do these solutions form a vector space? If so, what is its dimension?
- 7.4.23. Let  $L, M$  be linear functions. (a) Prove that  $\ker(L \circ M) \supseteq \ker M$ . (b) Find an example in which  $\ker(L \circ M) \neq \ker M$ .

## Inhomogeneous Systems

Now we turn our attention to inhomogeneous linear systems

$$L[\mathbf{u}] = \mathbf{f}, \quad (7.60)$$

where  $L: U \rightarrow V$  is a linear function,  $\mathbf{f} \in V$ , and the desired solution  $\mathbf{u} \in U$ . Unless  $\mathbf{f} = \mathbf{0}$ , the set of solutions to (7.60) is *not* a subspace of  $U$ , but, rather, forms an affine subspace, as defined in Exercise 2.2.28. Here, the crucial question is existence — is there a solution to the system? In contrast, for the homogeneous system  $L[\mathbf{z}] = \mathbf{0}$ , existence is not an issue, since  $\mathbf{0}$  is always a solution. The key question for homogeneous systems is uniqueness: either  $\ker L = \{\mathbf{0}\}$ , in which case  $\mathbf{0}$  is the only solution, or  $\ker L \neq \{\mathbf{0}\}$ , in which case there are infinitely many nontrivial solutions  $\mathbf{0} \neq \mathbf{z} \in \ker L$ .

In the matrix case, the compatibility of an inhomogeneous system  $A\mathbf{x} = \mathbf{b}$  — which was required for the existence of a solution — led to the general definition of the image of a matrix, which we copy verbatim for linear functions.

**Definition 7.37.** The *image* of a linear function  $L: U \rightarrow V$  is the subspace

$$\text{img } L = \{ L[\mathbf{u}] \mid \mathbf{u} \in U \} \subset V.$$

The proof that  $\text{img } L$  is a subspace of the codomain is straightforward: If  $\mathbf{f} = L[\mathbf{u}]$  and  $\mathbf{g} = L[\mathbf{v}]$  are any two elements of the image, so is any linear combination, since, by linearity

$$c\mathbf{f} + d\mathbf{g} = cL[\mathbf{u}] + dL[\mathbf{v}] = L[c\mathbf{u} + d\mathbf{v}] \in \text{img } L.$$

For example, if  $L[\mathbf{u}] = A\mathbf{u}$  is given by multiplication by an  $m \times n$  matrix, then its image is the subspace  $\text{img } L = \text{img } A \subset \mathbb{R}^m$  spanned by the columns of  $A$  — the *column space*

of the coefficient matrix. When  $L$  is a linear differential operator, or more general linear operator, characterizing its image can be a much more challenging problem.

The fundamental theorem regarding solutions to inhomogeneous linear equations exactly mimics our earlier result, Theorem 2.39, for matrix systems.

**Theorem 7.38.** Let  $L: U \rightarrow V$  be a linear function. Let  $\mathbf{f} \in V$ . Then the inhomogeneous linear system

$$L[\mathbf{u}] = \mathbf{f} \quad (7.61)$$

has a solution if and only if  $\mathbf{f} \in \text{img } L$ . In this case, the general solution to the system has the form

$$\mathbf{u} = \mathbf{u}^* + \mathbf{z}, \quad (7.62)$$

where  $\mathbf{u}^*$  is a particular solution, so  $L[\mathbf{u}^*] = \mathbf{f}$ , and  $\mathbf{z}$  is any element of  $\ker L$ , i.e., a solution to the corresponding homogeneous system

$$L[\mathbf{z}] = \mathbf{0}. \quad (7.63)$$

*Proof:* We merely repeat the proof of Theorem 2.39. The existence condition  $\mathbf{f} \in \text{img } L$  is an immediate consequence of the definition of the image. Suppose  $\mathbf{u}^*$  is a particular solution to (7.61). If  $\mathbf{z}$  is a solution to (7.63), then, by linearity,

$$L[\mathbf{u}^* + \mathbf{z}] = L[\mathbf{u}^*] + L[\mathbf{z}] = \mathbf{f} + \mathbf{0} = \mathbf{f},$$

and hence  $\mathbf{u}^* + \mathbf{z}$  is also a solution to (7.61). To show that every solution has this form, let  $\mathbf{u}$  be a second solution, so that  $L[\mathbf{u}] = \mathbf{f}$ . Setting  $\mathbf{z} = \mathbf{u} - \mathbf{u}^*$ , we find that

$$L[\mathbf{z}] = L[\mathbf{u} - \mathbf{u}^*] = L[\mathbf{u}] - L[\mathbf{u}^*] = \mathbf{f} - \mathbf{f} = \mathbf{0}.$$

Therefore  $\mathbf{z} \in \ker L$ , and so  $\mathbf{u}$  has the proper form (7.62). *Q.E.D.*

**Corollary 7.39.** The inhomogeneous linear system (7.61) has a *unique* solution if and only if  $\mathbf{f} \in \text{img } L$  and  $\ker L = \{\mathbf{0}\}$ .

Therefore, to prove that a linear system has a unique solution, we first need to prove an *existence result* that there is at least one solution, which requires the right-hand side  $\mathbf{f}$  to lie in the image of the operator  $L$ , and then a *uniqueness result*, that the only solution to the homogeneous system  $L[\mathbf{z}] = \mathbf{0}$  is the trivial zero solution  $\mathbf{z} = \mathbf{0}$ . Observe that whenever an inhomogeneous system  $L[\mathbf{u}] = \mathbf{f}$  has a unique solution, then every other inhomogeneous system  $L[\mathbf{u}] = \mathbf{g}$  that is defined by the *same* linear function also has a unique solution, provided  $\mathbf{g} \in \text{img } L$ . In other words, uniqueness does not depend upon the external forcing — although existence might.

**Remark.** In physical systems, the inhomogeneity  $\mathbf{f}$  typically corresponds to an external force. The decomposition formula (7.62) states that its effect on the linear system can be viewed as a combination of one specific response  $\mathbf{u}^*$  to the forcing and the system's internal, unencumbered motion, as represented by the homogeneous solution  $\mathbf{z}$ . Keep in mind that the particular solution is *not* uniquely defined (unless  $\ker L = \{\mathbf{0}\}$ ), and any one solution can serve in this role.

**Example 7.40.** Consider the inhomogeneous linear second order differential equation

$$u'' + u' - 2u = x. \quad (7.64)$$

Note that this can be written in the linear system form

$$L[u] = x, \quad \text{where} \quad L = D^2 + D - 2$$

is a linear second order differential operator. The kernel of the differential operator  $L$  is found by solving the associated homogeneous linear equation

$$L[z] = z'' + z' - 2z = 0. \quad (7.65)$$

Applying the usual solution method, we find that the homogeneous differential equation (7.65) has a two-dimensional solution space, with basis functions

$$z_1(x) = e^{-2x}, \quad z_2(x) = e^x.$$

Therefore, the general element of  $\ker L$  is a linear combination

$$z(x) = c_1 z_1(x) + c_2 z_2(x) = c_1 e^{-2x} + c_2 e^x.$$

To find a particular solution to the inhomogeneous differential equation (7.64), we rely on the method of *undetermined coefficients*<sup>†</sup>. We introduce the solution ansatz  $u = ax + b$ , and compute

$$L[u] = L[ax + b] = a - 2(ax + b) = -2ax + (a - 2b) = x.$$

Equating the coefficients of  $x$  and 1, and then solving for  $a = -\frac{1}{2}$ ,  $b = -\frac{1}{4}$ , we deduce that

$$u^*(x) = -\frac{1}{2}x - \frac{1}{4}$$

is a particular solution to the inhomogeneous differential equation. Theorem 7.38 then says that the general solution is

$$u(x) = u^*(x) + z(x) = -\frac{1}{2}x - \frac{1}{4} + c_1 e^{-2x} + c_2 e^x.$$

**Example 7.41.** By inspection, we see that

$$u(x, y) = -\frac{1}{2} \sin(x + y)$$

is a solution to the particular *Poisson equation*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \sin(x + y). \quad (7.66)$$

Theorem 7.38 implies that *every* solution to this inhomogeneous version of the Laplace equation (7.59) takes the form

$$u(x, y) = -\frac{1}{2} \sin(x + y) + z(x, y),$$

where  $z(x, y)$  is an arbitrary harmonic function, i.e., a solution to the homogeneous Laplace equation.

---

<sup>†</sup> One could also employ the method of *variation of parameters*, although usually the undetermined coefficient method, when applicable, is the more straightforward of the two. Details can be found in most ordinary differential equations texts, including [7, 22].

**Example 7.42.** Let us solve the second order linear boundary value problem

$$u'' + u = x, \quad u(0) = 0, \quad u(\pi) = 0. \quad (7.67)$$

As with initial value problems, the first step is to solve the differential equation. To this end, we first solve the corresponding homogeneous differential equation  $z'' + z = 0$ . The usual method — see [7] or Example 7.50 below — shows that  $\cos x$  and  $\sin x$  form a basis for its solution space. The method of undetermined coefficients then produces the particular solution  $u^*(x) = x$  to the inhomogeneous differential equation, and so its general solution is

$$u(x) = x + c_1 \cos x + c_2 \sin x. \quad (7.68)$$

The next step is to see whether any solutions also satisfy the boundary conditions. Plugging formula (7.68) into the boundary conditions yields

$$u(0) = c_1 = 0, \quad u(\pi) = \pi - c_1 = 0.$$

However, these two conditions are incompatible, and so there is *no* solution to the linear system (7.67). The function  $f(x) = x$  does not lie in the image of the differential operator  $L[u] = u'' + u$  when  $u$  is subjected to the boundary conditions. Or, to state it another way,  $(x, 0, 0)^T$  does not belong to the image of the linear operator  $M[u] = (u'' + u, u(0), u(\pi))^T$  defining the boundary value problem.

On the other hand, if we slightly modify the inhomogeneity, the boundary value problem

$$u'' + u = x - \frac{1}{2}\pi, \quad u(0) = 0, \quad u(\pi) = 0, \quad (7.69)$$

does admit a solution, but it fails to be unique. Applying the preceding solution techniques, we find that

$$u(x) = x - \frac{1}{2}\pi + \frac{1}{2}\pi \cos x + c \sin x$$

solves the system for *any* choice of constant  $c$ , and so the boundary value problem (7.69) admits infinitely many solutions. Observe that  $z(x) = \sin x$  is a basis for the kernel or solution space of the corresponding homogeneous boundary value problem

$$z'' + z = 0, \quad z(0) = 0, \quad z(\pi) = 0,$$

while  $u^*(x) = x - \frac{1}{2}\pi + \frac{1}{2}\pi \cos x$  represents a particular solution to the inhomogeneous system. Thus,  $u(x) = u^*(x) + z(x)$ , in conformity with the general formula (7.62).

Incidentally, if we modify the interval of definition, considering

$$u'' + u = f(x), \quad u(0) = 0, \quad u\left(\frac{1}{2}\pi\right) = 0, \quad (7.70)$$

then the homogeneous boundary value problem, with  $f(x) \equiv 0$ , has only the trivial solution, and so the inhomogeneous system admits a unique solution for *any* inhomogeneity  $f(x)$ . For example, if  $f(x) = x$ , then

$$u(x) = x - \frac{1}{2}\pi \sin x \quad (7.71)$$

is the unique solution to the resulting boundary value problem.

This example highlights some crucial differences between boundary value problems and initial value problems for ordinary differential equations. Nonsingular initial value problems have a unique solution for every suitable set of initial conditions. Boundary value problems have more of the flavor of linear algebraic systems, either possessing a unique solution for

all possible inhomogeneities, or admitting either no solution or infinitely many solutions, depending on the right-hand side. An interesting question is how to characterize the inhomogeneities  $f(x)$  that admit a solution, i.e., that lie in the image of the associated linear operator. These issues are explored in depth in [61].

## Exercises

7.4.24. For each of the following inhomogeneous systems, determine whether the right-hand side lies in the image of the coefficient matrix, and, if so, write out the general solution, clearly identifying the particular solution and the kernel element.

$$(a) \begin{pmatrix} 1 & -1 \\ 3 & -3 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & 1 & 4 \\ -1 & 2 & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 2 & -1 \\ 2 & 0 & 1 \\ 1 & -2 & 2 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix},$$

$$(d) \begin{pmatrix} -2 & 1 \\ -2 & 3 \\ 3 & -5 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad (e) \begin{pmatrix} -1 & 3 & 0 & 2 \\ 2 & -6 & 1 & -1 \\ -3 & 9 & -2 & 0 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 2 \\ -2 \\ 2 \end{pmatrix}.$$

7.4.25. Which of the following systems have a unique solution?

$$(a) \begin{pmatrix} 3 & 1 \\ -1 & -1 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 2 & -1 \\ -2 & 3 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

$$(c) \begin{pmatrix} 2 & 1 & -1 \\ 0 & -3 & -3 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} 3 \\ -1 \\ 5 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 4 & -1 \\ 1 & 3 & -3 \\ 2 & 3 & -2 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \\ 1 \end{pmatrix}.$$

7.4.26. Solve the following inhomogeneous linear ordinary differential equations:

$$(a) u' - 4u = x - 3, \quad (b) 5u'' - 4u' + 4u = e^x \cos x, \quad (c) u'' - 3u' = e^{3x}.$$

7.4.27. Solve the following initial value problems: (a)  $u' + 3u = e^x$ ,  $u(1) = 0$ , (b)  $u'' + 4u = 1$ ,  $u(\pi) = u'(\pi) = 0$ , (c)  $u'' - u' - 2u = e^x + e^{-x}$ ,  $u(0) = u'(0) = 0$ , (d)  $u'' + 2u' + 5u = \sin x$ ,  $u(0) = 1$ ,  $u'(0) = 0$ , (e)  $u''' - u'' + u' - u = x$ ,  $u(0) = 0$ ,  $u'(0) = 1$ ,  $u''(0) = 0$ .

7.4.28. Solve the following inhomogeneous Euler equations using either variation of parameters or the change of variables method discussed in Exercise 7.4.13:

$$(a) x^2u'' + xu' - u = x, \quad (b) x^2u'' - 2xu' + 2u = \log x, \quad (c) x^2u'' - 3xu' - 5u = 3x - 5.$$

7.4.29. Write down all solutions to the following boundary value problems. Label your answer as (i) unique solution, (ii) no solution, (iii) infinitely many solutions.

$$(a) u'' + 2u = 2x, \quad u(0) = 0, \quad u(\pi) = 0, \quad (b) u'' + 4u = \cos x, \quad u(-\pi) = 0, \quad u(\pi) = 1, \\ (c) u'' - 2u' + u = x - 2, \quad u(0) = -1, \quad u(1) = 1, \\ (d) u'' + 2u' + 2u = 1, \quad u(0) = \frac{1}{2}, \quad u(\pi) = \frac{1}{2}, \quad (e) u'' - 3u' + 2u = 4x, \quad u(0) = 0, \quad u(1) = 0, \\ (f) x^2u'' + xu' - u = 0, \quad u(0) = 1, \quad u(1) = 0, \quad (g) x^2u'' - 6u = 0, \quad u(1) = 1, \quad u(2) = -1, \\ (h) x^2u'' - 2xu' + 2u = 0, \quad u(0) = 0, \quad u(1) = 1.$$

◇ 7.4.30. Let  $L: U \rightarrow V$  be a linear function, and let  $W \subset U$  be a subspace of the domain space. (a) Prove that  $Y = \{L[\mathbf{w}] \mid \mathbf{w} \in W\} \subset \text{img } L \subset V$  is a subspace of the image. (b) Prove that  $\dim Y \leq \dim W$ . Conclude that a linear transformation can never increase the dimension of a subspace.

◇ 7.4.31. (a) Show that if  $L: V \rightarrow V$  is linear and  $\ker L \neq \{\mathbf{0}\}$ , then  $L$  is not invertible. (b) Show that if  $\text{img } L \neq V$ , then  $L$  is not invertible. (c) Give an example of a linear map with  $\ker L = \{\mathbf{0}\}$  that is not invertible. Hint: First explain why your example must be on an infinite-dimensional vector space.

## Superposition Principles for Inhomogeneous Systems

The *superposition principle* for inhomogeneous linear systems allows us to combine different inhomogeneities — provided that we do not change the underlying linear operator. The result is a straightforward generalization of the matrix version described in Theorem 2.44.

**Theorem 7.43.** Let  $L:U \rightarrow V$  be a linear function. Suppose that, for each  $i = 1, \dots, k$ , we know a particular solution  $\mathbf{u}_i^*$  to the inhomogeneous linear system  $L[\mathbf{u}] = \mathbf{f}_i$  for some  $\mathbf{f}_i \in \text{img } L$ . Then, given scalars  $c_1, \dots, c_k$ , a particular solution to the combined inhomogeneous system

$$L[\mathbf{u}] = c_1 \mathbf{f}_1 + \dots + c_k \mathbf{f}_k \quad (7.72)$$

is the corresponding linear combination

$$\mathbf{u}^* = c_1 \mathbf{u}_1^* + \dots + c_k \mathbf{u}_k^* \quad (7.73)$$

of particular solutions. The general solution to the inhomogeneous system (7.72) is

$$\mathbf{u} = \mathbf{u}^* + \mathbf{z} = c_1 \mathbf{u}_1^* + \dots + c_k \mathbf{u}_k^* + \mathbf{z}, \quad (7.74)$$

where  $\mathbf{z} \in \ker L$  is an arbitrary solution to the associated homogeneous system  $L[\mathbf{z}] = \mathbf{0}$ .

The proof is an easy consequence of linearity, and left to the reader. In physical terms, the superposition principle can be interpreted as follows. If we know the response of a linear physical system to several different external forces, represented by  $\mathbf{f}_1, \dots, \mathbf{f}_k$ , then the response of the system to a linear combination of these forces is just the self-same linear combination of the individual responses. The homogeneous solution  $\mathbf{z}$  represents an internal motion that the system acquires independent of any external forcing. Superposition relies on the linearity of the system, and so is always applicable in quantum mechanics, which is an inherently linear theory. On the other hand, in classical and relativistic mechanics, superposition is valid only in the linear approximation regime governing small motions/displacements/etc. Large-scale motions of a fully nonlinear physical system are more subtle, and combinations of external forces may lead to unexpected results.

**Example 7.44.** In Example 7.42, we found that a particular solution to the linear differential equation

$$u'' + u = x \quad \text{is} \quad u_1^* = x.$$

The method of undetermined coefficients can be used to solve the inhomogeneous equation

$$u'' + u = \cos x.$$

Since  $\cos x$  and  $\sin x$  are already solutions to the homogeneous equation, we must use the solution ansatz

$$u = ax \cos x + bx \sin x,$$

which, when substituted into the differential equation, produces the particular solution

$$u_2^* = -\frac{1}{2}x \sin x.$$

Therefore, by the superposition principle, the combined inhomogeneous system

$$u'' + u = 3x - 2 \cos x$$

has a particular solution

$$u^* = 3u_1^* - 2u_2^* = 3x + x \sin x.$$

The general solution is obtained by appending an arbitrary solution to the homogeneous equation:

$$u = 3x + x \sin x + c_1 \cos x + c_2 \sin x.$$

**Example 7.45.** Consider the boundary value problem

$$u'' + u = x, \quad u(0) = 2, \quad u\left(\frac{1}{2}\pi\right) = -1, \quad (7.75)$$

which is a modification of (7.70) with inhomogeneous boundary conditions. The superposition principle applies here, and allows us to decouple the inhomogeneity due to the forcing from the inhomogeneity due to the boundary conditions. We decompose the right-hand side, written in vectorial form, into simpler constituents<sup>†</sup>

$$\begin{pmatrix} x \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} x \\ 0 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

The first vector on the right-hand side corresponds to the preceding boundary value problem (7.70), whose solution was found in (7.71). The second and third vectors correspond to the unforced boundary value problems

$$u'' + u = 0, \quad u(0) = 1, \quad u\left(\frac{1}{2}\pi\right) = 0, \quad \text{and} \quad u'' + u = 0, \quad u(0) = 0, \quad u\left(\frac{1}{2}\pi\right) = 1,$$

with respective solutions  $u(x) = \cos x$  and  $u(x) = \sin x$ . Therefore, the solution to the combined boundary value problem (7.75) is the same linear combination of these individual solutions:

$$u(x) = \left(x - \frac{1}{2}\pi \sin x\right) + 2 \cos x - \sin x = x + 2 \cos x - \left(1 + \frac{1}{2}\pi\right) \sin x.$$

The solution is unique because the corresponding homogeneous boundary value problem

$$z'' + z = 0, \quad z(0) = 0, \quad z\left(\frac{1}{2}\pi\right) = 0,$$

has only the trivial solution  $z(x) \equiv 0$ , as you can verify.

## Exercises

7.4.32. Use superposition to solve the following inhomogeneous ordinary differential equations:

- (a)  $u' + 2u = 1 + \cos x$ , (b)  $u'' - 9u = x + \sin x$ , (c)  $9u'' - 18u' + 10u = 1 + e^x \cos x$ ,
- (d)  $u'' + u' - 2u = \sinh x$ , where  $\sinh x = \frac{1}{2}(e^x - e^{-x})$ , (e)  $u''' + 9u' = 1 + e^{3x}$ .

7.4.33. Consider the differential equation  $u'' + xu = 2$ . Suppose you know solutions to the two boundary value problems  $u(0) = 1$ ,  $u(1) = 0$  and  $u(0) = 0$ ,  $u(1) = 1$ . List all possible boundary value problems you can solve using superposition.

<sup>†</sup> **Warning.** When writing out a linear combination, make sure the scalars are *constants*! Writing the first summand as  $x(1, 0, 0)^T$  will lead to an *incorrect* application of the superposition principle.

7.4.34. Consider the differential equation  $xu'' - (x+1)u' + u = 0$ . Suppose we know the solution to the initial value problem  $u(1) = 2$ ,  $u'(1) = 1$  is  $u(x) = x+1$ , while the solution to the initial value problem  $u(1) = 1$ ,  $u'(1) = 1$  is  $u(x) = e^{x-1}$ . (a) What is the solution to the initial value problem  $u(1) = 3$ ,  $u'(1) = -2$ ? (b) What is the general solution to the differential equation?

7.4.35. Consider the differential equation  $4xu'' + 2u' + u = 0$ . Given that  $\cos\sqrt{x}$  solves the boundary value problem  $u\left(\frac{1}{4}\pi^2\right) = 0$ ,  $u(\pi^2) = -1$ , and  $\sin\sqrt{x}$  solves the boundary value problem  $u\left(\frac{1}{4}\pi^2\right) = 1$ ,  $u(\pi^2) = 0$ , write down the solution to the boundary value problem  $u\left(\frac{1}{4}\pi^2\right) = -3$ ,  $u(\pi^2) = 7$ .

7.4.36. Solve the following boundary value problems by using superposition: (a)  $u'' + 9u = x$ ,  $u(0) = 1$ ,  $u'(\pi) = 0$ , (b)  $u'' - 8u' + 16u = e^{4x}$ ,  $u(0) = 1$ ,  $u(1) = 0$ , (c)  $u'' + 4u = \sin 3x$ ,  $u'(0) = 0$ ,  $u(2\pi) = 3$ , (d)  $u'' - 2u' + u = 1 + e^x$ ,  $u'(0) = -1$ ,  $u'(1) = 1$ .

7.4.37. Given that  $x^2 + y^2$  solves the Poisson equation  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 4$ , while  $x^4 + y^4$  solves  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 12(x^2 + y^2)$ , write down a solution to  $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 1 + x^2 + y^2$ .

◇ 7.4.38. *Reduction of order:* Suppose you know one solution  $u_1(x)$  to the second order homogeneous differential equation  $u'' + a(x)u' + b(x)u = 0$ . (a) Show that if  $u(x) = v(x)u_1(x)$  is any other solution, then  $w(x) = v'(x)$  satisfies a first order differential equation. (b) Use reduction of order to find the general solution to the following equations, based on the indicated solution:

- (i)  $u'' - 2u' + u = 0$ ,  $u_1(x) = e^x$ , (ii)  $xu'' + (x-1)u' - u = 0$ ,  $u_1(x) = x-1$ ,
- (iii)  $u'' + 4xu' + (4x^2 + 2)u = 0$ ,  $u_1(x) = e^{-x^2}$ , (iv)  $u'' - (x^2 + 1)u = 0$ ,  $u_1(x) = e^{x^2/2}$ .

◇ 7.4.39. Write out the details of the proof of Theorem 7.43.

## Complex Solutions to Real Systems

As we know, solutions to a linear, homogeneous, constant coefficient ordinary differential equation are found by substituting an exponential ansatz, which effectively reduces the differential equation to the polynomial characteristic equation. Complex roots of the characteristic equation yield complex exponential solutions. But, if the equation is real, then the real and imaginary parts of the complex solutions are automatically real solutions. This solution technique is a particular case of a general principle for producing real solutions to real linear systems from, typically, simpler complex solutions. To work, the method requires us to impose some additional structure on the complex vector spaces involved.

**Definition 7.46.** A complex vector space  $V$  is called *conjugated* if it admits an operation of *complex conjugation* taking  $\mathbf{u} \in V$  to  $\bar{\mathbf{u}} \in V$  with the following properties:

- (a) conjugating twice returns one to the original vector:  $\bar{\bar{\mathbf{u}}} = \mathbf{u}$ ;
- (b) compatibility with vector addition:  $\bar{\mathbf{u} + v} = \bar{\mathbf{u}} + \bar{\mathbf{v}}$  for all  $\mathbf{u}, \mathbf{v} \in V$ ;
- (c) compatibility with scalar multiplication,  $\bar{\lambda\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}$ , for all  $\lambda \in \mathbb{C}$  and  $\mathbf{u} \in V$ .

The simplest example of a conjugated vector space is  $\mathbb{C}^n$ . The complex conjugate of a vector  $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$  is obtained by conjugating all its entries, whereby  $\bar{\mathbf{u}} = (\bar{u}_1, \bar{u}_2, \dots, \bar{u}_n)^T$ . Thus,

$$\begin{aligned} \mathbf{u} &= \mathbf{v} + i\mathbf{w}, & \text{where} & \quad \mathbf{v} = \operatorname{Re} \mathbf{u} = \frac{\mathbf{u} + \overline{\mathbf{u}}}{2}, & \quad \mathbf{w} = \operatorname{Im} \mathbf{u} = \frac{\mathbf{u} - \overline{\mathbf{u}}}{2i}, \end{aligned} \quad (7.76)$$

are the real and imaginary parts of  $\mathbf{u} \in \mathbb{C}^n$ . For example, if

$$\mathbf{u} = \begin{pmatrix} 1 - 2i \\ 3i \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix} + i \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix}, \quad \text{then} \quad \overline{\mathbf{u}} = \begin{pmatrix} 1 + 2i \\ -3i \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix} - i \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix},$$

whence

$$\operatorname{Re} \mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix}, \quad \operatorname{Im} \mathbf{u} = \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix}.$$

The other prototypical example of a conjugated vector space is the space of complex-valued functions  $f(x) = r(x) + is(x)$  defined on the interval  $a \leq x \leq b$ . The complex conjugate function is  $\overline{f}(x) = \overline{f(x)} = r(x) - is(x)$ . Thus, the complex conjugate of  $e^{(1+3i)x} = e^x \cos 3x + ie^x \sin 3x$  is  $\overline{e^{(1+3i)x}} = e^{(1-3i)x} = e^x \cos 3x - ie^x \sin 3x$ , with  $\operatorname{Re} e^{(1+3i)x} = e^x \cos 3x$ ,  $\operatorname{Im} e^{(1+3i)x} = ie^x \sin 3x$ .

An element  $\mathbf{v} \in V$  of a conjugated vector space is called *real* if  $\overline{\mathbf{v}} = \mathbf{v}$ . One easily checks that the real and imaginary parts of a general element, as defined by (7.76), are both real elements.

**Warning.** Not all subspaces of a conjugated vector space are conjugated. For example, the one-dimensional subspace of  $\mathbb{C}^2$  spanned by  $\mathbf{v}_1 = (1, 2)^T$  is conjugated. Indeed, the complex conjugate of a general element  $c\mathbf{v}_1 = (c, 2c)^T$  is  $(\overline{c}, 2\overline{c})^T = \overline{c}\mathbf{v}_1$  which also belongs to the subspace. On the other hand, the subspace spanned by  $(1, i)^T$  is *not* conjugated, because the complex conjugate of the element  $(c, ic)^T$  is  $(\overline{c}, -i\overline{c})^T$ , which does not belong to the subspace unless  $c = 0$ . In Exercise 7.4.50 you are asked to prove that a subspace  $V \subset \mathbb{C}^n$  is conjugated if and only if it has a basis  $\mathbf{v}_1, \dots, \mathbf{v}_k$  consisting entirely of real vectors. While conjugated subspaces play a role in certain applications, in practice we will deal only with  $\mathbb{C}^n$  and the entire space of complex-valued functions, and so can suppress most of these somewhat technical details.

**Definition 7.47.** A linear function  $L: U \rightarrow V$  between conjugated vector spaces is called *real* if it commutes with complex conjugation:

$$L[\overline{\mathbf{u}}] = \overline{L[\mathbf{u}]} \quad (7.77)$$

For example, any linear function  $L: \mathbb{C}^n \rightarrow \mathbb{C}^m$  is given by multiplication by an  $m \times n$  matrix:  $L[\mathbf{u}] = A\mathbf{u}$ . The function is real if and only if  $A$  is a real matrix. Similarly, a differential operator (7.15) is real if and only if its coefficients are real-valued functions.

The solutions to a homogeneous system defined by a real linear function satisfy the following general *Reality Principle*.

**Theorem 7.48.** If  $L[\mathbf{u}] = \mathbf{0}$  is a real homogeneous linear system and  $\mathbf{u} = \mathbf{v} + i\mathbf{w}$  is a complex solution, then its complex conjugate  $\overline{\mathbf{u}} = \mathbf{v} - i\mathbf{w}$  is also a solution. Moreover, both the real and imaginary parts,  $\mathbf{v}$  and  $\mathbf{w}$ , of a complex solution are real solutions.

*Proof:* First note that, by reality,  $L[\overline{\mathbf{u}}] = \overline{L[\mathbf{u}]} = \mathbf{0}$  whenever  $L[\mathbf{u}] = \mathbf{0}$ , and hence the complex conjugate  $\overline{\mathbf{u}}$  of any solution is also a solution. Therefore, by linear superposition,  $\mathbf{v} = \operatorname{Re} \mathbf{u} = \frac{1}{2}(\mathbf{u} + \overline{\mathbf{u}})$  and  $\mathbf{w} = \operatorname{Im} \mathbf{u} = \frac{1}{2i}(\mathbf{u} - \overline{\mathbf{u}})$  are also solutions. *Q.E.D.*

**Example 7.49.** The real linear matrix system

$$\begin{pmatrix} 2 & -1 & 3 & 0 \\ -2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has a complex solution

$$\mathbf{u} = \begin{pmatrix} -1 - 3i \\ 1 \\ 1 + 2i \\ -2 - 4i \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -2 \end{pmatrix} + i \begin{pmatrix} -3 \\ 0 \\ 2 \\ -4 \end{pmatrix}.$$

Since the coefficient matrix is real, the real and imaginary parts,

$$\mathbf{v} = (-1, 1, 1, -2)^T, \quad \mathbf{w} = (-3, 0, 2, -4)^T,$$

are both solutions of the system, as can easily be checked.

On the other hand, the complex linear system

$$\begin{pmatrix} 2 & -2i & i & 0 \\ 1+i & 0 & -2-i & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has the complex solution

$$\mathbf{u} = \begin{pmatrix} 1-i \\ -i \\ 2 \\ 2+2i \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \end{pmatrix} + i \begin{pmatrix} -1 \\ -1 \\ 0 \\ 2 \end{pmatrix}.$$

However, neither its real nor its imaginary part is a solution to the system.

**Example 7.50.** Consider the real homogeneous ordinary differential equation

$$u'' + 2u' + 5u = 0.$$

To solve it, we use the usual exponential ansatz  $u = e^{\lambda x}$ , leading to the characteristic equation

$$\lambda^2 + 2\lambda + 5 = 0.$$

There are two roots,

$$\lambda_1 = -1 + 2i, \quad \lambda_2 = -1 - 2i,$$

leading, via Euler's formula (3.92), to the complex solutions

$$\begin{aligned} u_1(x) &= e^{(-1+2i)x} = e^{-x} \cos 2x + i e^{-x} \sin 2x, \\ u_2(x) &= e^{(-1-2i)x} = e^{-x} \cos 2x - i e^{-x} \sin 2x. \end{aligned}$$

The complex conjugate of the first solution is the second, in accordance with Theorem 7.48. Moreover, the real and imaginary parts of the two solutions

$$v(x) = e^{-x} \cos 2x, \quad w(x) = e^{-x} \sin 2x,$$

are individual real solutions. Since the solution space is two-dimensional, the general solution is a linear combination

$$u(x) = c_1 e^{-x} \cos 2x + c_2 e^{-x} \sin 2x,$$

of the two linearly independent real solutions.

**Example 7.51.** Consider the real second order Euler differential equation

$$L[u] = x^2 u'' + 7x u' + 13u = 0.$$

The roots of the associated characteristic equation

$$r(r-1) + 7r + 13 = r^2 + 6r + 13 = 0$$

are complex:  $r = -3 \pm 2i$ , and the resulting solutions  $x^{-3+2i}, x^{-3-2i}$  are complex conjugate powers. We use Euler's formula (3.92), to obtain their real and imaginary parts:

$$x^{-3+2i} = x^{-3} e^{2i \log x} = x^{-3} \cos(2 \log x) + i x^{-3} \sin(2 \log x),$$

valid for  $x > 0$ . (For  $x < 0$  just replace  $x$  by  $-x$  in the above formula.) Again, by Theorem 7.48, the real and imaginary parts of the complex solution are by themselves real solutions to the equation. Therefore, the general real solution to this differential equation for  $x > 0$  is

$$u(x) = c_1 x^{-3} \cos(2 \log x) + c_2 x^{-3} \sin(2 \log x).$$

**Example 7.52.** The complex monomial

$$u(x, y) = (x + iy)^n$$

is a solution to the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

because, by the chain rule,

$$\frac{\partial^2 u}{\partial x^2} = n(n-1)(x+iy)^{n-2}, \quad \frac{\partial^2 u}{\partial y^2} = n(n-1)i^2(x+iy)^{n-2} = -n(n-1)(x+iy)^{n-2}.$$

Since the Laplacian operator is real, Theorem 7.48 implies that the real and imaginary parts of this complex solution are real solutions. The resulting real solutions are the harmonic polynomials introduced in Example 7.36.

Knowing this, it is relatively easy to find the explicit formulas for the harmonic polynomials. We appeal to the binomial formula

$$(a+b)^n = \sum_{i=0}^n \binom{n}{k} a^k b^{n-k}, \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (7.78)$$

is the standard notation for the *binomial coefficients*. Since  $i^2 = -1$ ,  $i^3 = -i$ ,  $i^4 = 1$ , etc., we have

$$\begin{aligned} (x+iy)^n &= x^n + nx^{n-1}(iy) + \binom{n}{2} x^{n-2}(iy)^2 + \binom{n}{3} x^{n-3}(iy)^3 + \cdots + (iy)^n \\ &= x^n + i n x^{n-1} y - \binom{n}{2} x^{n-2} y^2 - i \binom{n}{3} x^{n-3} y^3 + \cdots. \end{aligned}$$

Separating the real and imaginary terms, we obtain the explicit formulas

$$\begin{aligned} \operatorname{Re} (x+iy)^n &= x^n - \binom{n}{2} x^{n-2} y^2 + \binom{n}{4} x^{n-4} y^4 + \cdots, \\ \operatorname{Im} (x+iy)^n &= n x^{n-1} y - \binom{n}{3} x^{n-3} y^3 + \binom{n}{5} x^{n-5} y^5 + \cdots, \end{aligned} \quad (7.79)$$

for the two independent harmonic polynomials of degree  $n$ . The first few of these polynomials were described in Example 7.36. In fact, it can be proved that the most general solution to the Laplace equation can be written as a convergent infinite series in the basic harmonic polynomials, cf. [61].

## Exercises

7.4.40. Can you find a complex matrix  $A$  such that  $\ker A \neq \{\mathbf{0}\}$  and the real and imaginary parts of every complex solution to  $A\mathbf{u} = \mathbf{0}$  are also solutions?

7.4.41. Find the general real solution to the following homogeneous differential equations:

- (a)  $u'' + 4u = 0$ , (b)  $u'' + 6u' + 10u = 0$ , (c)  $2u''' + 3u' - 5u = 0$ , (d)  $u'''' + u = 0$ ,  
 (e)  $u'''' + 13u'' + 36u = 0$ , (f)  $x^2u'' - xu' + 3u = 0$ , (g)  $x^3u''' + x^2u'' + 3xu' - 8u = 0$ .

7.4.42. The following functions are solutions to a real constant coefficient homogeneous scalar ordinary differential equation. (i) Determine the least possible order of the differential equation, and (ii) write down an appropriate differential equation. (a)  $e^{-x} \sin 3x$ ,  
 (b)  $x \sin x$ , (c)  $1 + xe^{-x} \cos 2x$ , (d)  $\sin x + \cos 2x$ , (e)  $\sin x + x^2 \cos x$ .

7.4.43. Find the general solution to the following complex ordinary differential equations.

Verify that, in these cases, the real and imaginary parts of a complex solution are *not* real solutions. (a)  $u' + iu = 0$ , (b)  $u'' - iu' + (i-1)u = 0$ , (c)  $u'' - iu = 0$ .

7.4.44. (a) Write down the explicit formulas for the harmonic polynomials of degree 4 and check that they are indeed solutions to the Laplace equation. (b) Prove that every homogeneous polynomial solution of degree 4 is a linear combination of the two basic harmonic polynomials.

7.4.45. Find all complex exponential solutions  $u(t, x) = e^{\omega t + kx}$  of the *beam equation*  $\frac{\partial^2 u}{\partial t^2} = \frac{\partial^4 u}{\partial x^4}$ . How many different real solutions can you produce?

◇ 7.4.46. (a) Show that, if  $k \in \mathbb{R}$ , then  $u(t, x) = e^{-k^2 t + ikx}$  is a complex solution to the *heat equation*  $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ . (b) Use complex conjugation to write down another complex solution. (c) Find two independent real solutions to the heat equation. (d) Can  $k$  be complex? If so, what real solutions are produced? (e) Which of your solutions decay to zero as  $t \rightarrow \infty$ ? (f) Can you solve the exercise assuming  $k \in \mathbb{C} \setminus \mathbb{R}$  is not real?

7.4.47. Show that the free space *Schrödinger equation*  $i \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$  is not a real linear system by constructing a complex quadratic polynomial solution and verifying that its real and imaginary parts are not solutions.

7.4.48. Which of the following sets of vectors span conjugated subspaces of  $\mathbb{C}^3$ ?

- (a)  $\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ ; (b)  $\begin{pmatrix} 1 \\ -i \\ 2i \end{pmatrix}$ ; (c)  $\begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$ ; (d)  $\begin{pmatrix} 1 \\ 0 \\ i \end{pmatrix}$ ,  $\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}$ ; (e)  $\begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ 0 \\ -i \end{pmatrix}$ ,  $\begin{pmatrix} 0 \\ 1 \\ i \end{pmatrix}$ .

◇ 7.4.49. Prove that the real and imaginary parts of a general element of a conjugated vector space, as defined by (7.76), are both real elements.

◇ 7.4.50. Prove that a subspace  $V \subset \mathbb{C}^n$  is conjugated if and only if it admits a basis all of whose elements are real.

◇ 7.4.51. Prove that if  $L[\mathbf{u}] = \mathbf{f}$  is a real inhomogeneous linear system with real right-hand side  $\mathbf{f}$ , and  $\mathbf{u} = \mathbf{v} + i\mathbf{w}$  is a complex solution, then its real part  $\mathbf{v}$  is a solution to the system,  $L[\mathbf{v}] = \mathbf{f}$ , while its imaginary part  $\mathbf{w}$  solves the homogeneous system  $L[\mathbf{w}] = \mathbf{0}$ .

- ◇ 7.4.52. Prove that a linear function  $L: \mathbb{C}^n \rightarrow \mathbb{C}^m$  is real if and only if  $L[\mathbf{u}] = A\mathbf{u}$ , where  $A$  is a real  $m \times n$  matrix.
- ◇ 7.4.53. Let  $\mathbf{u} = \mathbf{x} + i\mathbf{y}$  be a complex solution to a real linear system. Under what conditions are its real and imaginary parts  $\mathbf{x}, \mathbf{y}$  linearly independent real solutions?

---

## 7.5 Adjoints, Positive Definite Operators, and Minimization Principles

Sections 2.5 and 4.4 revealed the importance of the adjoint system  $A^T \mathbf{y} = \mathbf{f}$  in the analysis of systems of linear algebraic equations  $A\mathbf{x} = \mathbf{b}$ . Two of the four fundamental matrix subspaces are based on the transposed matrix. While the  $m \times n$  matrix  $A$  defines a linear function from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , its transpose,  $A^T$ , has size  $n \times m$  and hence characterizes a linear function in the *reverse* direction, from  $\mathbb{R}^m$  to  $\mathbb{R}^n$ .

As with most basic concepts for linear algebraic systems, the adjoint system and transpose operation on the coefficient matrix are the prototypes of a more general construction that is valid for general linear functions. However, it is not immediately obvious how to “transpose” a more general linear operator  $L[u]$ , e.g., a differential operator acting on function space. In this section, we shall introduce the abstract concept of the *adjoint* of a linear function that generalizes the transpose operation on matrices. This will be followed by a general characterization of positive definite linear operators and the characterization of the solutions to the associated linear systems via minimization principles. Unfortunately, we will not have sufficient analytical tools to develop most of the interesting examples, and instead refer the interested reader to [61, 79].

The adjoint (and transpose) rely on an inner product structure on both the domain and codomain spaces. For simplicity, we restrict our attention to real inner product spaces, leaving the complex version to the interested reader. Thus, we begin with a linear function  $L: U \rightarrow V$  that maps an inner product space  $U$  to a second inner product space  $V$ . We distinguish the inner products on  $U$  and  $V$  (which may be different even when  $U$  and  $V$  are the same vector space) by using a single angle bracket

$$\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle \quad \text{to denote the inner product between} \quad \mathbf{u}, \tilde{\mathbf{u}} \in U,$$

and a double angle bracket

$$\langle\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle\rangle \quad \text{to denote the inner product between} \quad \mathbf{v}, \tilde{\mathbf{v}} \in V.$$

Once inner products on both the domain and codomain are prescribed, the abstract definition of the adjoint of a linear function can be formulated.

**Definition 7.53.** Let  $U, V$  be inner product spaces, and let  $L: U \rightarrow V$  be a linear function. The *adjoint* of  $L$  is the function<sup>†</sup>  $L^*: V \rightarrow U$  that satisfies

$$\langle\langle L[\mathbf{u}], \mathbf{v} \rangle\rangle = \langle \mathbf{u}, L^*[\mathbf{v}] \rangle \quad \text{for all} \quad \mathbf{u} \in U, \quad \mathbf{v} \in V. \quad (7.80)$$

---

<sup>†</sup> The notation  $L^*$  unfortunately coincides with that of the dual linear function introduced in Exercise 7.2.30. These clashing notations are both well established in the literature, although occasionally a prime, as in  $V', L'$ , is used for dual spaces, maps, etc. However, it is possible to reconcile the two notations in a natural manner; see Exercise 7.5.10. In this book, the dual notation appears only in these few exercises.

Note that the adjoint function goes in the *opposite* direction to  $L$ , just like the transposed matrix. Also, the left-hand side of equation (7.80) indicates the inner product on  $V$ , while the right-hand side is the inner product on  $U$  — which is where the respective vectors live. In infinite-dimensional situations, the adjoint may not exist. But if it does, then it is uniquely determined by (7.80); see Exercise 7.5.7.

**Remark.** Technically, (7.80) serves to define the “formal adjoint” of the linear operator  $L$ . For the infinite-dimensional function spaces arising in analysis, a true adjoint must satisfy certain additional analytical requirements, [50, 67]. However, for pedagogical reasons, it is better to suppress such advanced analytical complications in this introductory treatment.

**Lemma 7.54.** The adjoint of a linear function is a linear function.

*Proof:* Given  $\mathbf{u} \in U$ ,  $\mathbf{v}, \mathbf{w} \in V$ , and scalars  $c, d \in \mathbb{R}$ , using the defining property of the adjoint and the bilinearity of the two inner products produces

$$\begin{aligned}\langle \mathbf{u}, L^*[c\mathbf{v} + d\mathbf{w}] \rangle &= \langle L[\mathbf{u}], c\mathbf{v} + d\mathbf{w} \rangle = c \langle L[\mathbf{u}], \mathbf{v} \rangle + d \langle L[\mathbf{u}], \mathbf{w} \rangle \\ &= c \langle \mathbf{u}, L^*[\mathbf{v}] \rangle + d \langle \mathbf{u}, L^*[\mathbf{w}] \rangle = \langle \mathbf{u}, cL^*[\mathbf{v}] + dL^*[\mathbf{w}] \rangle.\end{aligned}$$

Since this holds for all  $\mathbf{u} \in U$ , we must have

$$L^*[c\mathbf{v} + d\mathbf{w}] = cL^*[\mathbf{v}] + dL^*[\mathbf{w}], \quad \text{thereby proving linearity.} \quad Q.E.D.$$

**Example 7.55.** Let us first show how the defining equation (7.80) for the adjoint leads directly to the transpose of a matrix. Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  be the linear function  $L[\mathbf{v}] = A\mathbf{v}$  defined by multiplication by the  $m \times n$  matrix  $A$ . Then  $L^*: \mathbb{R}^m \rightarrow \mathbb{R}^n$  is linear, and so is represented by matrix multiplication,  $L^*[\mathbf{v}] = A^*\mathbf{v}$ , by an  $n \times m$  matrix  $A^*$ . We impose the ordinary Euclidean dot products

$$\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle = \mathbf{u} \cdot \tilde{\mathbf{u}} = \mathbf{u}^T \tilde{\mathbf{u}}, \quad \mathbf{u}, \tilde{\mathbf{u}} \in \mathbb{R}^n, \quad \langle \mathbf{v}, \tilde{\mathbf{v}} \rangle = \mathbf{v} \cdot \tilde{\mathbf{v}} = \mathbf{v}^T \tilde{\mathbf{v}}, \quad \mathbf{v}, \tilde{\mathbf{v}} \in \mathbb{R}^m,$$

as our inner products on both  $\mathbb{R}^n$  and  $\mathbb{R}^m$ . Evaluation of both sides of the adjoint identity (7.80) yields

$$\begin{aligned}\langle L[\mathbf{u}], \mathbf{v} \rangle &= \langle A\mathbf{u}, \mathbf{v} \rangle = (A\mathbf{u})^T \mathbf{v} = \mathbf{u}^T A^T \mathbf{v}, \\ \langle \mathbf{u}, L^*[\mathbf{v}] \rangle &= \langle \mathbf{u}, A^* \mathbf{v} \rangle = \mathbf{u}^T A^* \mathbf{v}.\end{aligned}\tag{7.81}$$

Since these expressions must agree for all  $\mathbf{u}, \mathbf{v}$ , the matrix  $A^*$  representing  $L^*$  is equal to the transposed matrix  $A^T$ , as justified in Exercise 1.6.13. We conclude that *the adjoint of a matrix with respect to the Euclidean dot product is its transpose*:  $A^* = A^T$ .

**Remark.** See Exercise 7.2.30 for another interpretation of the transpose in terms of dual vector spaces. Again, keep in mind that the  $^*$  notation has a different meaning there.

**Example 7.56.** Let us now adopt different, weighted inner products on the domain and codomain for the linear map  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  given by  $L[\mathbf{u}] = A\mathbf{v}$ . Suppose that

- (i) the inner product on the domain space  $\mathbb{R}^n$  is given by  $\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle = \mathbf{u}^T M \tilde{\mathbf{u}}$ , while
- (ii) the inner product on the codomain  $\mathbb{R}^m$  is given by  $\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle = \mathbf{v}^T C \tilde{\mathbf{v}}$ .

Here  $M$  and  $C$  are positive definite matrices of respective sizes  $n \times n$  and  $m \times m$ . Then, in place of (7.81), we have

$$\langle A\mathbf{u}, \mathbf{v} \rangle = (A\mathbf{u})^T C \mathbf{v} = \mathbf{u}^T A^T C \mathbf{v}, \quad \langle \mathbf{u}, A^* \mathbf{v} \rangle = \mathbf{u}^T M A^* \mathbf{v}.$$

Equating these expressions, we deduce that  $A^T C = M A^*$ . Therefore, the *weighted adjoint* of the matrix  $A$  is given by the more complicated formula

$$A^* = M^{-1} A^T C. \quad (7.82)$$

In mechanical applications,  $M$  plays the role of the mass matrix, and explicitly appears in the dynamical systems to be studied in Chapter 10. In particular, suppose  $A$  is square, defining a linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$ . If we adopt the same inner product  $\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle = \mathbf{v}^T C \tilde{\mathbf{v}}$  on both the domain and codomain spaces  $\mathbb{R}^n$ , then its adjoint matrix  $A^* = C^{-1} A^T C$  is similar to its transpose.

Everything that we learned about transposes can be reinterpreted in the more general language of adjoints. First, applying the adjoint operation twice returns you to where you began; this is an immediate consequence of the defining equation (7.80).

**Proposition 7.57.** The adjoint of the adjoint of  $L$  is just  $L = (L^*)^*$ .

The next result generalizes the fact, (1.55), that the transpose of the product of two matrices is the product of the transposes, in the reverse order.

**Proposition 7.58.** Let  $U, V, W$  be inner product spaces. If  $L: U \rightarrow V$  and  $M: V \rightarrow W$  have respective adjoints  $L^*: V \rightarrow U$  and  $M^*: W \rightarrow V$ , then the composite linear function  $M \circ L: U \rightarrow W$  has adjoint  $(M \circ L)^* = L^* \circ M^*$ , which maps  $W$  to  $U$ .

*Proof:* Let  $\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle, \langle \mathbf{v}, \tilde{\mathbf{v}} \rangle, \langle \mathbf{w}, \tilde{\mathbf{w}} \rangle$ , denote, respectively, the inner products on  $U, V, W$ . For  $\mathbf{u} \in U, \mathbf{w} \in W$ , we compute using the definition (7.80) repeatedly:

$$\begin{aligned} \langle \mathbf{u}, (M \circ L)^*[\mathbf{w}] \rangle &= \langle \mathbf{u}, M[L[\mathbf{u}]] \rangle = \langle M[L[\mathbf{u}]], \mathbf{w} \rangle \\ &= \langle L[\mathbf{u}], M^*[\mathbf{w}] \rangle = \langle \mathbf{u}, L^*[M^*[\mathbf{w}]] \rangle = \langle \mathbf{u}, L^* \circ M^*[\mathbf{w}] \rangle. \end{aligned}$$

Since this holds for all  $\mathbf{u}$  and  $\mathbf{w}$ , the identification follows. Q.E.D.

In this chapter, we have been able to actually compute adjoints in just the finite-dimensional situation, when the linear functions are given by matrix multiplication. For the more challenging case of adjoints of linear operators on function spaces, e.g., differential operators appearing in boundary value problems, the reader should consult [61].

## Exercises

7.5.1. Choose one from the following list of inner products on  $\mathbb{R}^2$ . Then find the adjoint of

$A = \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}$  when your inner product is used on both its domain and codomain. (a) the Euclidean dot product; (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1 w_1 + 3v_2 w_2$ ; (c) the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T K \mathbf{w}$  defined by the positive definite matrix  $K = \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}$ .

7.5.2. From the list in Exercise 7.5.1, choose different inner products on the domain and codomain, and then determine the adjoint of the matrix  $A$ .

7.5.3. Choose one from the following list of inner products on  $\mathbb{R}^3$  for both the domain and

codomain, and find the adjoint of  $A = \begin{pmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 2 \end{pmatrix}$ : (a) the Euclidean dot product;

- (b) the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + 2v_2 w_2 + 3v_3 w_3$ ; (c) the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T K \mathbf{w}$  defined by the positive definite matrix  $K = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ .

7.5.4. From the list in Exercise 7.5.3, choose different inner products on the domain and codomain, and then compute the adjoint of the matrix  $A$ .

7.5.5. Choose an inner product on  $\mathbb{R}^2$  from the list in Exercise 7.5.1, and an inner product on  $\mathbb{R}^3$  from the list in Exercise 7.5.3, and then compute the adjoint of  $A = \begin{pmatrix} 1 & 3 \\ 0 & 2 \\ -1 & 1 \end{pmatrix}$ .

7.5.6. Let  $\mathcal{P}^{(2)}$  be the space of quadratic polynomials equipped with the inner product  $\langle p, q \rangle = \int_0^1 p(x) q(x) dx$ . Find the adjoint of the derivative operator  $D[p] = p'$  on  $\mathcal{P}^{(2)}$ .

◇ 7.5.7. Prove that, if it exists, the adjoint of a linear function is uniquely determined by (7.80).

◇ 7.5.8. Prove that (a)  $(L + M)^* = L^* + M^*$ , (b)  $(cL)^* = cL^*$  for  $c \in \mathbb{R}$ ,  
(c)  $(L^*)^* = L$ , (d)  $(L^{-1})^* = (L^*)^{-1}$ .

◇ 7.5.9. Let  $L: U \rightarrow V$  be a linear function between inner product spaces. Prove that  $\mathbf{u} \in \mathbb{R}^n$  solves the inhomogeneous linear system  $L[\mathbf{u}] = \mathbf{f}$  if and only if

$$\langle \mathbf{u}, L^*[\mathbf{v}] \rangle = \langle \mathbf{f}, \mathbf{v} \rangle \quad \text{for all } \mathbf{v} \in V. \quad (7.83)$$

Explain why Exercise 3.1.11 is a special case of this result. **Remark.** Equation (7.83) is known as the *weak formulation* of the linear system. It plays an essential role in the analysis of differential equations and their numerical approximations, [61].

◇ 7.5.10. Suppose  $V, W$  are finite-dimensional inner product spaces with dual space  $V^*, W^*$ . Let  $L: V \rightarrow W$  be a linear function, and let  $\tilde{L}^*: W^* \rightarrow V^*$  denote the dual linear function, as in Exercise 7.2.30 (without the tilde), while  $L^*: W \rightarrow V$  denotes its adjoint. (As noted above, the same notation denotes two mathematically different objects.) Prove that if we identify  $V^* \simeq V$  and  $W^* \simeq W$  using the linear isomorphism in Exercise 7.1.62, then the dual and adjoint functions are identified  $\tilde{L}^* \simeq L^*$ , thus reconciling the unfortunate clash in notation. In particular, this includes the two possible interpretations of the transpose of a matrix.

## Self-Adjoint and Positive Definite Linear Functions

Throughout this section  $U$  will be an inner product space. We will show how to generalize the notions of symmetric and positive definite matrices to linear operators on  $U$  in a natural fashion. First, we define the analogue of a symmetric matrix.

**Definition 7.59.** A linear function  $J: U \rightarrow U$  is called *self-adjoint* if  $J^* = J$ . A self-adjoint linear function is *positive definite*, written  $J > 0$ , if

$$\langle \mathbf{u}, J[\mathbf{u}] \rangle > 0 \quad \text{for all } \mathbf{0} \neq \mathbf{u} \in U. \quad (7.84)$$

In particular, if  $J > 0$  then  $\ker J = \{\mathbf{0}\}$ . (Why?) Thus, a positive definite linear system  $J[\mathbf{u}] = \mathbf{f}$  with  $\mathbf{f} \in \text{img } J$  must have a unique solution. The next result generalizes our basic observation that the Gram matrices  $A^T A$  and  $A^T C A$ , cf. (3.62, 64), are symmetric and positive (semi-)definite.

**Theorem 7.60.** Let  $L:U \rightarrow V$  be a linear map between inner product spaces with adjoint  $L^*:V \rightarrow U$ . Then the composite map  $J = L^* \circ L:U \rightarrow U$  is self-adjoint. Moreover,  $J$  is positive definite if and only if  $\ker L = \{\mathbf{0}\}$ .

*Proof:* First, by Propositions 7.58 and 7.57,

$$J^* = (L^* \circ L)^* = L^* \circ (L^*)^* = L^* \circ L = J,$$

proving self-adjointness. Furthermore, for  $\mathbf{u} \in U$ , the inner product

$$\langle \mathbf{u}, J[\mathbf{u}] \rangle = \langle \mathbf{u}, L^*[L[\mathbf{u}]] \rangle = \langle L[\mathbf{u}], L[\mathbf{u}] \rangle = \|L[\mathbf{u}]\|^2 > 0$$

is strictly positive, provided that  $L[\mathbf{u}] \neq \mathbf{0}$ . Thus, if  $\ker L = \{\mathbf{0}\}$ , then the positivity condition (7.84) holds, and conversely. *Q.E.D.*

Let us specialize to the case of a linear function  $L:\mathbb{R}^n \rightarrow \mathbb{R}^m$  that is represented by the  $m \times n$  matrix  $A$ , so that  $L[\mathbf{u}] = A\mathbf{u}$ . When the Euclidean dot product is used on the two spaces, the adjoint  $L^*$  is represented by the transpose  $A^T$ , and hence the map  $J = L^* \circ L$  has matrix representation  $J[\mathbf{u}] = K\mathbf{u}$ , where  $K = A^T A$ . Therefore, in this case Theorem 7.60 reduces to our earlier Proposition 3.36, governing the positive definiteness of the Gram matrix product  $A^T A$ . If we change the inner product on the codomain to  $\langle \mathbf{w}, \tilde{\mathbf{w}} \rangle = \mathbf{w}^T C \tilde{\mathbf{w}}$  for some  $C > 0$ , then  $L^*$  is represented by  $A^T C$ , and hence  $J = L^* \circ L$  has matrix form  $K = A^T C A$ , which is the general symmetric, positive definite Gram matrix constructed in (3.64) that underlay our development of the equations of equilibrium in Chapter 6.

Finally, if we further replace the dot product on the domain space  $\mathbb{R}^n$  by the alternative inner product  $\langle \mathbf{v}, \tilde{\mathbf{v}} \rangle = \mathbf{v}^T M \tilde{\mathbf{v}}$  for  $M > 0$ , then, according to formula (7.82), the adjoint of  $L$  has matrix form

$$A^* = M^{-1} A^T C, \quad \text{and therefore} \quad K = A^* A = M^{-1} A^T C A \quad (7.85)$$

is a self-adjoint, positive (semi-)definite matrix with respect to the weighted inner product on  $\mathbb{R}^n$  prescribed by the positive definite matrix  $M$ . In this case, the positive definite, self-adjoint operator  $J$  is *no longer represented by a symmetric matrix*. So, we did not quite tell the truth when we said we would allow only symmetric matrices to be positive definite — we really meant only self-adjoint matrices.

General self-adjoint matrices will be important in our discussion of the vibrations of mass–spring chains that have unequal masses. Extensions of these constructions to differential operators underlies the analysis of the boundary value problems of continuum mechanics, to be studied in [61].

## Exercises

7.5.11. Show that the following linear transformations of  $\mathbb{R}^2$  are self-adjoint with respect to the Euclidean dot product: (a) rotation through the angle  $\theta = \pi$ ; (b) reflection about the line  $y = x$ . (c) The scaling map  $S[\mathbf{x}] = 3\mathbf{x}$ ; (d) orthogonal projection onto the line  $y = x$ .

◇ 7.5.12. Let  $M$  be a positive definite matrix. Show that  $A:\mathbb{R}^n \rightarrow \mathbb{R}^n$  is self-adjoint with respect to the inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T M \mathbf{w}$  if and only if  $MA$  is a symmetric matrix.

7.5.13. Prove that  $A = \begin{pmatrix} 6 & 3 \\ 2 & 4 \end{pmatrix}$  is self-adjoint with respect to the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = 2v_1 w_1 + 3v_2 w_2$ . *Hint:* Use the criterion in Exercise 7.5.12.

7.5.14. Consider the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = v_1 w_1 + \frac{1}{2}v_2 w_2 + \frac{1}{3}v_3 w_3$  on  $\mathbb{R}^3$ .

(a) What are the conditions on the entries of a  $3 \times 3$  matrix  $A$  in order that it be self-adjoint? *Hint:* Use the criterion in Exercise 7.5.12. (b) Write down an example of a non-diagonal self-adjoint matrix.

7.5.15. Answer Exercise 7.5.14 for the inner product based on  $\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ .

7.5.16. *True or false:* The identity transformation is self-adjoint for an arbitrary inner product on the underlying vector space.

7.5.17. *True or false:* A diagonal matrix is self-adjoint for an arbitrary inner product on  $\mathbb{R}^n$ .

7.5.18. Suppose  $L: U \rightarrow U$  has an adjoint  $L^*: U \rightarrow U$ . (a) Show that  $L + L^*$  is self-adjoint.

(b) Show that  $L \circ L^*$  is self-adjoint.

◇ 7.5.19. Suppose  $J, M: U \rightarrow U$  are self-adjoint linear functions on an inner product space  $U$ .

(a) Prove that  $\langle J[\mathbf{u}], \mathbf{u} \rangle = \langle M[\mathbf{u}], \mathbf{u} \rangle$  for all  $\mathbf{u} \in U$  if and only if  $J = M$ .

(b) Explain why this result is false if the self-adjointness hypothesis is dropped.

7.5.20. Prove that if  $L: U \rightarrow U$  is an invertible linear transformation on an inner product space  $U$ , then the following three statements are equivalent: (a)  $\langle L[\mathbf{u}], L[\mathbf{v}] \rangle = \langle \mathbf{u}, \mathbf{v} \rangle$  for all  $\mathbf{u}, \mathbf{v} \in U$ . (b)  $\|L[\mathbf{u}]\| = \|\mathbf{u}\|$  for all  $\mathbf{u} \in U$ . (c)  $L^* = L^{-1}$ . *Hint:* Use Exercise 7.5.19.

7.5.21. (a) Prove that the operation  $M_a[u(x)] = a(x)u(x)$  of multiplication by a continuous function  $a(x)$  defines a self-adjoint linear operator on the function space  $C^0[a, b]$  with respect to the  $L^2$  inner product. (b) Is  $M_a$  also self-adjoint with respect to the weighted inner product  $\langle\langle f, g \rangle\rangle = \int_a^b f(x)g(x)w(x)dx$ ?

◇ 7.5.22. A linear function  $S: U \rightarrow U$  is called *skew-adjoint* if  $S^* = -S$ . (a) Prove that a skew-symmetric matrix is skew-adjoint with respect to the standard dot product on  $\mathbb{R}^n$ . (b) Under what conditions is  $S[\mathbf{x}] = A\mathbf{x}$  skew-adjoint with respect to the inner product  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T M \mathbf{y}$  on  $\mathbb{R}^n$ ? (c) Let  $L: U \rightarrow U$  have an adjoint  $L^*$ . Prove that  $L - L^*$  is skew-adjoint. (d) Explain why every linear operator  $L: U \rightarrow U$  that has an adjoint  $L^*$  can be written as the sum of a self-adjoint and a skew-adjoint operator.

◇ 7.5.23. (a) Let  $L_1: U \rightarrow V_1$  and  $L_2: U \rightarrow V_2$  be linear maps between inner product spaces, with  $V_1, V_2$  not necessarily the same. Let  $J_1 = L_1^* \circ L_1$ ,  $J_2 = L_2^* \circ L_2$ . Show that the sum  $J = J_1 + J_2$  can be written as a self-adjoint combination  $J = L^* \circ L$  for some linear operator  $L$ . *Hint:* See Exercise 3.4.35 for the matrix case.

## Minimization

In Chapter 5, we learned how the solution to a linear algebraic system  $K\mathbf{u} = \mathbf{f}$  with positive definite coefficient matrix  $K$  can be characterized as the unique minimizer for the quadratic function  $p(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T K \mathbf{u} - \mathbf{u}^T \mathbf{f}$ . There is an analogous minimization principle that characterizes the solutions to linear systems defined by positive definite linear operators. This general result is of tremendous importance in analysis of boundary value problems for differential equations, for both physical and mathematical reasons, and also inspires the finite element numerical solution algorithm, [61].

We restrict our attention to real linear functions on real vector spaces in this section.

**Theorem 7.61.** Let  $J: U \rightarrow U$  be a positive definite linear function on a real inner product space  $U$ . If  $\mathbf{f} \in \text{img } J$ , then the quadratic function

$$p(\mathbf{u}) = \frac{1}{2} \langle \mathbf{u}, J[\mathbf{u}] \rangle - \langle \mathbf{u}, \mathbf{f} \rangle \quad (7.86)$$

has a unique minimizer  $\mathbf{u} = \mathbf{u}^*$ , which is the solution to the linear system  $J[\mathbf{u}] = \mathbf{f}$ .

*Proof:* The proof mimics that of its matrix counterpart in Theorem 5.2. Our assumption that  $\mathbf{f} \in \text{img } J$  implies that there is a  $\mathbf{u}^* \in U$  such that  $J[\mathbf{u}^*] = \mathbf{f}$ . Thus, we can write

$$p(\mathbf{u}) = \frac{1}{2} \langle \mathbf{u}, J[\mathbf{u}] \rangle - \langle \mathbf{u}, J[\mathbf{u}^*] \rangle = \frac{1}{2} \langle \mathbf{u} - \mathbf{u}^*, J[\mathbf{u} - \mathbf{u}^*] \rangle - \frac{1}{2} \langle \mathbf{u}^*, J[\mathbf{u}^*] \rangle, \quad (7.87)$$

where we used linearity, along with the fact that  $J$  is self-adjoint to identify the terms  $\langle \mathbf{u}, J[\mathbf{u}^*] \rangle = \langle \mathbf{u}^*, J[\mathbf{u}] \rangle$ . Since  $J > 0$ , the first term on the right-hand side of (7.87) is always  $\geq 0$ ; moreover, it equals its minimal value 0 if and only if  $\mathbf{u} = \mathbf{u}^*$ . On the other hand, the second term does not depend upon  $\mathbf{u}$  at all, and hence is unaffected by variations in  $\mathbf{u}$ . Therefore, to minimize  $p(\mathbf{u})$ , we must make the first term as small as possible, which is accomplished by setting  $\mathbf{u} = \mathbf{u}^*$ . *Q.E.D.*

**Remark.** For linear functions given by matrix multiplication, positive definiteness automatically implies invertibility, and so the linear system  $K\mathbf{u} = \mathbf{f}$  has a solution for every right-hand side. This is not so immediate when  $J$  is a positive definite operator on an infinite-dimensional function space. Therefore, the existence of a solution or minimizer is a significant issue. And, in fact, many modern analytical existence results rely on the determination of suitable minimization principles. On the other hand, once existence is assured, uniqueness follows immediately from the positive definiteness of the operator  $J$ .

**Theorem 7.62.** Suppose  $L: U \rightarrow V$  is a linear function between inner product spaces with  $\ker L = \{\mathbf{0}\}$  and adjoint function  $L^*: V \rightarrow U$ . Let  $J = L^* \circ L: U \rightarrow U$  be the associated positive definite linear function. If  $\mathbf{f} \in \text{img } J$ , then the quadratic function

$$p(\mathbf{u}) = \frac{1}{2} \|L[\mathbf{u}]\|^2 - \langle \mathbf{u}, \mathbf{f} \rangle \quad (7.88)$$

has a unique minimizer  $\mathbf{u}^*$ , which is the solution to the linear system  $J[\mathbf{u}^*] = \mathbf{f}$ .

*Proof:* It suffices to note that the quadratic term in (7.88) can be written in the alternative form

$$\|L[\mathbf{u}]\|^2 = \langle L[\mathbf{u}], L[\mathbf{u}] \rangle = \langle \mathbf{u}, L^*[L[\mathbf{u}]] \rangle = \langle \mathbf{u}, J[\mathbf{u}] \rangle.$$

Thus, (7.88) reduces to the quadratic function of the form (7.86) with  $J = L^* \circ L$ , and so Theorem 7.62 follows directly from Theorem 7.61. *Q.E.D.*

**Warning.** In (7.88), the first term  $\|L[\mathbf{u}]\|^2$  is computed using the norm based on the inner product on  $V$ , while the second term  $\langle \mathbf{u}, \mathbf{f} \rangle$  employs the inner product on  $U$ .

**Example 7.63.** For a general positive definite matrix (7.85), the quadratic function (7.88) is computed with respect to the alternative inner product  $\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle = \mathbf{u}^T M \tilde{\mathbf{u}}$ , so

$$p(\mathbf{u}) = \frac{1}{2} \|A\mathbf{u}\|^2 - \langle \mathbf{u}, \mathbf{f} \rangle = \frac{1}{2} (A\mathbf{u})^T C A \mathbf{u} - \mathbf{u}^T M \mathbf{f} = \frac{1}{2} \mathbf{u}^T (A^T C A) \mathbf{u} - \mathbf{u}^T (M\mathbf{f}).$$

Theorem 7.62 tells us that the minimizer of the quadratic function is the solution to

$$A^T C A \mathbf{u} = M\mathbf{f}, \quad \text{which we rewrite as} \quad K\mathbf{u} = M^{-1} A^T C A \mathbf{u} = \mathbf{f}.$$

This conclusion also follows from our earlier finite-dimensional Minimization Theorem 5.2.

In [61, 79], it is shown that the most important minimization principles that characterize solutions to the linear boundary value problems of physics and engineering all arise through this remarkably general mathematical construction.

## Exercises

7.5.24. Find the minimum value of  $p(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T \begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix} \mathbf{u} - \mathbf{u}^T \begin{pmatrix} 1 \\ -1 \end{pmatrix}$  for  $\mathbf{u} \in \mathbb{R}^2$ .

7.5.25. Minimize the function  $p(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T \begin{pmatrix} 2 & -1 & 0 \\ -1 & 4 & -2 \\ 0 & -2 & 3 \end{pmatrix} \mathbf{u} - \mathbf{u}^T \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}$  for  $\mathbf{u} \in \mathbb{R}^3$ .

7.5.26. Minimize  $\|(2x - y, x + y)^T\|^2 - 6x$  over all  $x, y$ , where  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^2$ .

7.5.27. Answer Exercise 7.5.26 for (a) the weighted norm  $\|(x, y)^T\| = \sqrt{2x^2 + 3y^2}$ ;

(b) the norm based on  $\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ ; (c) the norm based on  $\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ .

7.5.28. Let  $L(x, y) = \begin{pmatrix} x - 2y \\ x + y \\ -x + 3y \end{pmatrix}$  and  $\mathbf{f} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ . Minimize  $p(\mathbf{x}) = \frac{1}{2} \|L[\mathbf{x}]\|^2 - \langle \mathbf{x}, \mathbf{f} \rangle$  using

(a) the Euclidean inner products and norms on both  $\mathbb{R}^2$  and  $\mathbb{R}^3$ ; (b) the Euclidean inner product on  $\mathbb{R}^2$  and the weighted norm  $\|\mathbf{w}\| = \sqrt{w_1^2 + 2w_2^2 + 3w_3^2}$  on  $\mathbb{R}^3$ ; (c) the inner product given by  $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$  on  $\mathbb{R}^2$  and the Euclidean norm on  $\mathbb{R}^3$ ; (d) the inner product given by  $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$  on  $\mathbb{R}^2$  and the weighted norm  $\|\mathbf{w}\| = \sqrt{w_1^2 + 2w_2^2 + 3w_3^2}$  on  $\mathbb{R}^3$ .

7.5.29. Find the minimum distance between the point  $(1, 0, 0)^T$  and the plane  $x + y - z = 0$  when distance is measured in (a) the Euclidean norm; (b) the weighted norm  $\|\mathbf{w}\| =$

$\sqrt{w_1^2 + 2w_2^2 + 3w_3^2}$ ; (c) the norm based on the positive definite matrix  $\begin{pmatrix} 3 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 3 \end{pmatrix}$ .

◇ 7.5.30. How would you modify the statement of Theorem 7.62 if  $\ker L \neq \{\mathbf{0}\}$ ?



# Chapter 8

# Eigenvalues and Singular Values

So far, our physical applications of linear algebra have concentrated on statics: unchanging equilibrium configurations of mass–spring chains, circuits, and structures, all modeled by linear systems of algebraic equations. It is now time to set the universe in motion. In general, a (continuous) *dynamical system* refers to the (differential) equations governing the time-varying motion of a physical system, be it mechanical, electrical, chemical, fluid, thermodynamical, biological, financial, . . . . Our immediate goal is to solve the simplest class of continuous dynamical models, which are first order autonomous linear systems of ordinary differential equations.

We begin with a very quick review of the scalar case, whose solutions are simple exponential functions. This inspires us to try to solve a vector-valued linear system by substituting a similar exponential solution formula. We are immediately led to the system of algebraic equations that define the eigenvalues and eigenvectors of the coefficient matrix. Thus, before we can make any progress in our study of differential equations, we need to learn about eigenvalues and eigenvectors, and that is the purpose of the present chapter. Dynamical systems are used to motivate the subject, but serious applications will be deferred until Chapter 10. Additional applications of eigenvalues and eigenvectors to linear iterative systems, stochastic processes, and numerical solution algorithms for linear algebraic systems form the focus of Chapter 9.

Each square matrix possesses a collection of one or more complex scalars, called eigenvalues, and associated vectors, called eigenvectors. From a geometrical viewpoint, the matrix defines a linear transformation on Euclidean space; the eigenvectors indicate the directions of pure stretch and the eigenvalues the extent of stretching. We will introduce the non-standard term “complete” to describe matrices whose (complex) eigenvectors form a basis of the underlying vector space. A more common name for such matrices is “diagonalizable”, because, when expressed in terms of its eigenvector basis, the matrix representing the corresponding linear transformation assumes a very simple diagonal form, facilitating the detailed analysis of its properties. A particularly important class consists of the symmetric matrices, whose eigenvectors form an orthogonal basis of  $\mathbb{R}^n$ ; in fact, this is how orthogonal bases most naturally appear. Most matrices are complete; incomplete matrices are trickier to deal with, and we discuss their non-diagonal Schur decomposition and Jordan canonical form in Section 8.6.

A non-square matrix  $A$  does not possess eigenvalues. In their place, one studies the eigenvalues of the associated square Gram matrix  $K = A^T A$ , whose square roots are known as the singular values of the original matrix. The corresponding singular value decomposition (SVD) supplies the final details for our understanding of the remarkable geometric structure governing matrix multiplication. The singular value decomposition is used to define the pseudoinverse of a matrix, which provides a mechanism for “inverting” non-square and singular matrices, and an alternative construction of least squares solutions to general linear systems. Singular values underlie the powerful method of modern statistical data analysis known as principal component analysis (PCA), which will be developed in the

final section of this chapter and appears in an increasingly broad range of contemporary applications, including image processing, semantics, language and speech recognition, and machine learning.

**Remark.** The numerical computation of eigenvalues and eigenvectors is a challenging issue, and must be deferred until Section 9.5. Unless you are prepared to consult that section in advance, solving the computer-based problems in this chapter will require access to computer software that can accurately compute eigenvalues and eigenvectors.

## 8.1 Linear Dynamical Systems

Our new goal is to solve and analyze the simplest class of dynamical systems, namely those modeled by first order linear systems of ordinary differential equations. We begin with a thorough review of the scalar case, including a complete investigation into the stability of their equilibria — in preparation for the general situation to be treated in depth in Chapter 10. Readers who are not interested in such motivational material may skip ahead to Section 8.2 without incurring any penalty.

### Scalar Ordinary Differential Equations

Consider the elementary first order scalar ordinary differential equation

$$\frac{du}{dt} = a u. \quad (8.1)$$

Here  $a \in \mathbb{R}$  is a real constant, while the unknown  $u(t)$  is a scalar function. As you no doubt already learned, e.g., in [7, 22, 78], the general solution to (8.1) is an exponential function

$$u(t) = c e^{at}. \quad (8.2)$$

The integration constant  $c$  is uniquely determined by a single *initial condition*

$$u(t_0) = b \quad (8.3)$$

imposed at an initial time  $t_0$ . Substituting  $t = t_0$  into the solution formula (8.2),

$$u(t_0) = c e^{at_0} = b, \quad \text{and so} \quad c = b e^{-at_0}.$$

We conclude that

$$u(t) = b e^{a(t-t_0)} \quad (8.4)$$

is the unique solution to the scalar initial value problem (8.1, 3).

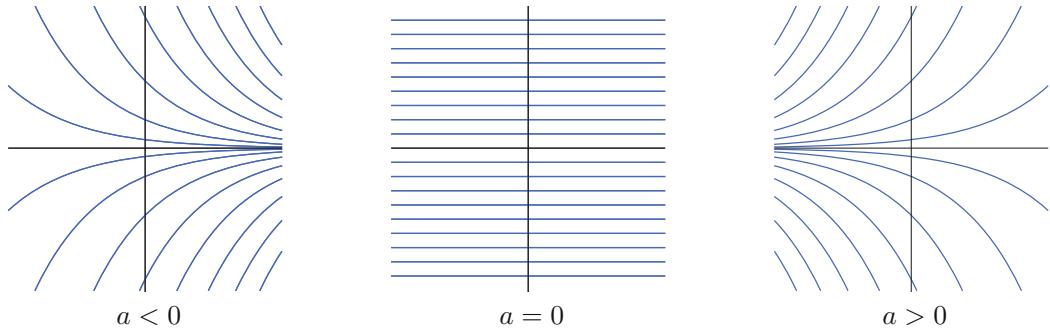
**Example 8.1.** The radioactive decay of an isotope, say uranium-238, is governed by the differential equation

$$\frac{du}{dt} = -\gamma u. \quad (8.5)$$

Here  $u(t)$  denotes the amount of the isotope remaining at time  $t$ ; the coefficient  $\gamma > 0$  governs the decay rate. The solution is an exponentially decaying function  $u(t) = c e^{-\gamma t}$ , where  $c = u(0)$  is the amount of radioactive material at the initial time  $t_0 = 0$ .

The isotope's *half-life*  $t_*$  is the time it takes for half of a sample to decay, that is, when  $u(t_*) = \frac{1}{2}u(0)$ . To determine  $t_*$ , we solve the algebraic equation

$$e^{-\gamma t_*} = \frac{1}{2}, \quad \text{so that} \quad t_* = \frac{\log 2}{\gamma}. \quad (8.6)$$



**Figure 8.1.** Solutions to  $\dot{u} = a u$ .

Thus, after an elapsed time of  $t_*$ , half of the original material has decayed. After a further  $t_*$ , half of the remaining half has decayed, leaving a quarter of the original radioactive material. And so on, so that at each integer multiple  $nt_*$ ,  $n \in \mathbb{N}$ , of the half-life, the remaining amount of the isotope is  $u(nt_*) = 2^{-n} u(0)$ .

Returning to the general situation, let us make some elementary, but pertinent, observations about this simplest linear dynamical system. First of all, since the equation is homogeneous, the zero function  $u(t) \equiv 0$  — corresponding to  $c = 0$  in the solution formula (8.2) — is a constant solution, known as an *equilibrium solution*, or *fixed point*, since it does not depend on  $t$ . If the coefficient  $a > 0$  is positive, then the solutions (8.2) are exponentially growing (in absolute value) as  $t \rightarrow +\infty$ . This implies that the zero equilibrium solution is *unstable*. The initial condition  $u(t_0) = 0$  produces the zero solution, but if we make a tiny error (either physical, numerical, or mathematical) in the initial data, say  $u(t_0) = \varepsilon$ , then the solution  $u(t) = \varepsilon e^{a(t-t_0)}$  will eventually be far away from equilibrium. More generally, any two solutions with very close, but not equal, initial data will eventually become arbitrarily far apart:  $|u_1(t) - u_2(t)| \rightarrow \infty$  as  $t \rightarrow \infty$ . One consequence is an inherent difficulty in accurately computing the long-time behavior of the solution, since small numerical errors may eventually have very large effects.

On the other hand, if  $a < 0$ , the solutions are exponentially decaying in time. In this case, the zero equilibrium solution is *stable*, since a small change in the initial data will have a negligible effect on the solution. In fact, the zero solution is *globally asymptotically stable*. The phrase “asymptotically stable” indicates that solutions that start out near the equilibrium point approach it in the large time limit; more specifically, if  $u(t_0) = \varepsilon$  is small, then  $u(t) \rightarrow 0$  as  $t \rightarrow \infty$ . “Globally” implies that *all* solutions, no matter how large the initial data, eventually approach equilibrium. In fact, for a linear system, the stability of an equilibrium solution is inevitably a global phenomenon.

The borderline case is  $a = 0$ . Then all the solutions to (8.1) are constant. In this case, the zero solution is *stable* — indeed, *globally stable* — but not asymptotically stable. The solution to the initial value problem  $u(t_0) = \varepsilon$  is  $u(t) \equiv \varepsilon$ . Therefore, a solution that starts out near equilibrium will remain nearby, but will not asymptotically approach it. The three qualitatively different possibilities are illustrated in Figure 8.1. A formal definition of the various notions of stability for dynamical systems can be found in Definition 10.14.

## Exercises

8.1.1. Solve the following initial value problems:

(a)  $\frac{du}{dt} = 5u$ ,  $u(0) = -3$ , (b)  $\frac{du}{dt} = 2u$ ,  $u(1) = 3$ , (c)  $\frac{du}{dt} = -3u$ ,  $u(-1) = 1$ .

8.1.2. Suppose a radioactive material has a half-life of 100 years. What is the decay rate  $\gamma$ ?

Starting with an initial sample of 100 grams, how much will be left after 10 years?  
after 100 years? after 1,000 years?

8.1.3. Carbon-14 has a half-life of 5730 years. Human skeletal fragments discovered in a cave are analyzed and found to have only 6.24% of the carbon-14 that living tissue would have. How old are the remains?

8.1.4. Prove that if  $t_*$  is the half-life of a radioactive material, then  $u(nt_*) = 2^{-n}u(0)$ . Explain the meaning of this equation in your own words.

8.1.5. A bacteria colony grows according to the equation  $du/dt = 1.3u$ . How long until the colony doubles? quadruples? If the initial population is 2, how long until the population reaches 2 million?

8.1.6. Deer in northern Minnesota reproduce according to the linear differential equation  $\frac{du}{dt} = .27u$  where  $t$  is measured in years. If the initial population is  $u(0) = 5,000$  and the environment can sustain at most 1,000,000 deer, how long until the deer run out of resources?

◇ 8.1.7. Consider the inhomogeneous differential equation  $\frac{du}{dt} = au + b$ , where  $a, b$  are constants.

- (a) Show that  $u_* = -b/a$  is a constant equilibrium solution. (b) Solve the differential equation. *Hint:* Look at the differential equation satisfied by  $v = u - u_*$ .  
(c) Discuss the stability of the equilibrium solution  $u_*$ .

8.1.8. Use the method of Exercise 8.1.7 to solve the following initial value problems:

(a)  $\frac{du}{dt} = 2u - 1$ ,  $u(0) = 1$ , (b)  $\frac{du}{dt} = 5u + 15$ ,  $u(1) = -3$ , (c)  $\frac{du}{dt} = -3u + 6$ ,  $u(2) = -1$ .

8.1.9. The radioactive waste from a nuclear reactor has a half-life of 1000 years. Waste is continually produced at the rate of 5 tons per year and stored in a dump site.

- (a) Set up an inhomogeneous differential equation, of the form in Exercise 8.1.7, to model the amount of radioactive waste. (b) Determine whether the amount of radioactive material at the dump increases indefinitely, decreases to zero, or eventually stabilizes at some fixed amount. (c) Starting with a brand new site, how long will it be until the dump contains 100 tons of radioactive material?

◇ 8.1.10. Suppose that hunters are allowed to shoot a fixed number of the northern Minnesota deer in Exercise 8.1.6 each year. (a) Explain why the population model takes the form  $\frac{du}{dt} = .27u - b$ , where  $b$  is the number killed yearly. (Ignore the seasonal aspects of hunting.)

- (b) If  $b = 1,000$ , how long until the deer run out of resources? *Hint:* See Exercise 8.1.7.  
(c) What is the maximal rate at which deer can be hunted without causing their extinction?

8.1.11. (a) Write down the exact solution to the initial value problem  $\frac{du}{dt} = \frac{2}{7}u$ ,  $u(0) = \frac{1}{3}$ .

- (b) Suppose you make the approximation  $u(0) = .3333$ . At what point does your solution differ from the true solution by 1 unit? by 1000 units? (c) Answer the same question if you also approximate the coefficient in the differential equation by  $\frac{du}{dt} = .2857u$ .

- ◇ 8.1.12. Let  $a$  be complex. Prove that  $u(t) = ce^{at}$  is the (complex) solution to our scalar ordinary differential equation (8.1). Describe the asymptotic behavior of the solution as  $t \rightarrow \infty$ , and the stability properties of the zero equilibrium solution.
- ◇ 8.1.13. (a) Prove that if  $u_1(t)$  and  $u_2(t)$  are any two distinct solutions to  $\frac{du}{dt} = au$  with  $a > 0$ , then  $|u_1(t) - u_2(t)| \rightarrow \infty$  as  $t \rightarrow \infty$ . (b) If  $a = .02$  and  $u_1(0) = .1$ ,  $u_2(0) = .05$ , how long do you have to wait until  $|u_1(t) - u_2(t)| > 1,000$ ?

## First Order Dynamical Systems

The simplest class of *dynamical system* is a coupled system of  $n$  first order ordinary differential equations

$$\frac{du_1}{dt} = f_1(t, u_1, \dots, u_n), \quad \dots \quad \frac{du_n}{dt} = f_n(t, u_1, \dots, u_n).$$

The unknowns are the  $n$  scalar functions  $u_1(t), \dots, u_n(t)$  depending on the scalar variable  $t \in \mathbb{R}$ , which we usually view as time, whence the term “dynamical”. We will often write the system in the equivalent vector form

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}), \tag{8.7}$$

in which  $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))^T$  is the vector-valued solution, which serves to parameterize a curve in  $\mathbb{R}^n$ , and  $\mathbf{f}(t, \mathbf{u})$  is a vector-valued function with components  $f_i(t, u_1, \dots, u_n)$  for  $i = 1, \dots, n$ . A dynamical system is called *autonomous* if the time variable  $t$  does not appear explicitly on the right-hand side, and so has the form

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}). \tag{8.8}$$

Dynamical systems of ordinary differential equations appear in an astonishing variety of applications, including physics, astronomy, chemistry, biology, weather and climate, economics and finance, and have been intensely studied since the very first days following the invention of calculus.

In this text, we shall concentrate most of our attention on the very simplest case: a homogeneous, linear, autonomous dynamical system, in which the right-hand side  $f(\mathbf{u})$  is a linear function of  $\mathbf{u}$  that is independent of the time  $t$ , and hence given by multiplication by a constant matrix. Thus, the system takes the form

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}, \tag{8.9}$$

in which  $A$  is a constant  $n \times n$  matrix. Writing out the system (8.9) in full detail produces

$$\begin{aligned} \frac{du_1}{dt} &= a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n, \\ \frac{du_2}{dt} &= a_{21}u_1 + a_{22}u_2 + \dots + a_{2n}u_n, \\ &\vdots && \vdots \\ \frac{du_n}{dt} &= a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nn}u_n, \end{aligned} \tag{8.10}$$

and we seek the solution  $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))^T$ . In the autonomous case, which is the only type to be treated in depth here, the coefficients  $a_{ij}$  are assumed to be (real) constants. We are interested not only in the formulas for the solutions, but also in understanding their qualitative and quantitative behavior.

Drawing our inspiration from the exponential solution formula (8.2) for the scalar version, let us investigate whether the vector system admits any solutions of a similar exponential form

$$\mathbf{u}(t) = e^{\lambda t} \mathbf{v}. \quad (8.11)$$

We assume that  $\lambda$  is a constant scalar, so  $e^{\lambda t}$  is the usual scalar exponential function, while  $\mathbf{v} \in \mathbb{R}^n$  is a constant vector. In other words, the components  $u_i(t) = e^{\lambda t} v_i$  of our desired solution are assumed to be constant multiples of the *same* exponential function. Since  $\mathbf{v}$  is constant, the derivative of  $\mathbf{u}(t)$  is easily found:

$$\frac{d\mathbf{u}}{dt} = \frac{d}{dt} (e^{\lambda t} \mathbf{v}) = \lambda e^{\lambda t} \mathbf{v}.$$

On the other hand, since  $e^{\lambda t}$  is a scalar, it commutes with matrix multiplication, and so

$$A\mathbf{u} = A e^{\lambda t} \mathbf{v} = e^{\lambda t} A\mathbf{v}.$$

Therefore,  $\mathbf{u}(t)$  will solve the system (8.9) if and only if

$$\lambda e^{\lambda t} \mathbf{v} = e^{\lambda t} A\mathbf{v},$$

or, canceling the common scalar factor  $e^{\lambda t}$ ,

$$\lambda \mathbf{v} = A\mathbf{v}.$$

The result is a system of  $n$  algebraic equations relating the vector  $\mathbf{v}$  and the scalar  $\lambda$ . Analysis of this system and its ramifications will be the topic of the remainder of this chapter. A broad range of significant applications will appear in the subsequent two chapters.

## 8.2 Eigenvalues and Eigenvectors

In view of the preceding motivational section, we hereby inaugurate our discussion of eigenvalues and eigenvectors by stating the basic definition.

**Definition 8.2.** Let  $A$  be an  $n \times n$  matrix. A scalar  $\lambda$  is called an *eigenvalue* of  $A$  if there is a *non-zero* vector  $\mathbf{v} \neq \mathbf{0}$ , called an *eigenvector*, such that

$$A\mathbf{v} = \lambda \mathbf{v}. \quad (8.12)$$

In geometric terms, the matrix  $A$  has the effect of stretching the eigenvector  $\mathbf{v}$  by an amount specified by the eigenvalue  $\lambda$ .

**Remark.** The odd-looking terms “eigenvalue” and “eigenvector” are hybrid German–English words. In the original German, they are *Eigenwert* and *Eigenvektor*, which can be fully translated as “proper value” and “proper vector”. For some reason, the half-translated terms have acquired a certain charm, and are now standard. The alternative English terms *characteristic value* and *characteristic vector* can be found in some (mostly older) texts. Oddly, the terms *characteristic polynomial* and *characteristic equation*, to be defined below, are still used rather than “eigenpolynomial” and “eigenequation”.

The requirement that the eigenvector  $\mathbf{v}$  be nonzero is important, since  $\mathbf{v} = \mathbf{0}$  is a trivial solution to the eigenvalue equation (8.12) for *every* scalar  $\lambda$ . Moreover, as far as solving

linear ordinary differential equations goes, the zero vector  $\mathbf{v} = \mathbf{0}$  gives  $\mathbf{u}(t) \equiv \mathbf{0}$ , which is certainly a solution, but one that we already knew.

The eigenvalue equation (8.12) is a system of linear equations for the entries of the eigenvector  $\mathbf{v}$  — provided that the eigenvalue  $\lambda$  is specified in advance — but is “mildly” nonlinear as a combined system for  $\lambda$  and  $\mathbf{v}$ . Gaussian Elimination per se will not solve the problem, and we are in need of a new idea. Let us begin by rewriting the equation in the form<sup>†</sup>

$$(A - \lambda I)\mathbf{v} = \mathbf{0}, \quad (8.13)$$

where  $I$  is the identity matrix of the correct size, whereby  $\lambda I\mathbf{v} = \lambda\mathbf{v}$ . Now, for given  $\lambda$ , equation (8.13) is a homogeneous linear system for  $\mathbf{v}$ , and always has the trivial zero solution  $\mathbf{v} = \mathbf{0}$ . But we are specifically seeking a nonzero solution! According to Theorem 1.47, a homogeneous linear system has a nonzero solution  $\mathbf{v} \neq \mathbf{0}$  if and only if its coefficient matrix, which in this case is  $A - \lambda I$ , is singular. This observation is the key to resolving the eigenvector equation.

**Theorem 8.3.** A scalar  $\lambda$  is an eigenvalue of the  $n \times n$  matrix  $A$  if and only if the matrix  $A - \lambda I$  is singular, i.e., of rank  $< n$ . The corresponding eigenvectors are the nonzero solutions to the eigenvalue equation  $(A - \lambda I)\mathbf{v} = \mathbf{0}$ .

We know a number of ways to characterize singular matrices, including the vanishing determinant criterion given in (1.84). Therefore, the following result is immediate:

**Corollary 8.4.** A scalar  $\lambda$  is an eigenvalue of the matrix  $A$  if and only if  $\lambda$  is a solution to the *characteristic equation*

$$\det(A - \lambda I) = 0. \quad (8.14)$$

In practice, when computing eigenvalues and eigenvectors by hand using exact arithmetic, one first solves the characteristic equation (8.14) to obtain the set of eigenvalues. Then, for each eigenvalue  $\lambda$  one uses standard linear algebra methods, e.g., Gaussian Elimination, to solve the corresponding linear system (8.13) for the associated eigenvector  $\mathbf{v}$ .

**Example 8.5.** Consider the  $2 \times 2$  matrix

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

We compute the determinant in the characteristic equation using formula (1.38):

$$\det(A - \lambda I) = \det \begin{pmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{pmatrix} = (3 - \lambda)^2 - 1 = \lambda^2 - 6\lambda + 8.$$

Thus, the characteristic equation is a quadratic polynomial equation, and can be solved by factorization:

$$\lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2) = 0.$$

We conclude that  $A$  has two eigenvalues:  $\lambda_1 = 4$  and  $\lambda_2 = 2$ .

For each eigenvalue, the corresponding eigenvectors are found by solving the associated homogeneous linear system (8.13). For the first eigenvalue, the eigenvector equation is

$$(A - 4I)\mathbf{v} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{or} \quad \begin{aligned} -x + y &= 0, \\ x - y &= 0. \end{aligned}$$

---

<sup>†</sup> Note that it is not legal to write (8.13) in the form  $(A - \lambda)\mathbf{v} = \mathbf{0}$  since we do not know how to subtract a scalar  $\lambda$  from a matrix  $A$ . Worse, if you type  $A - \lambda$  in MATLAB or MATHEMATICA, the result will be to subtract  $\lambda$  from *all* the entries of  $A$ , which is *not* what we are after!

The general solution is

$$x = y = a, \quad \text{so} \quad \mathbf{v} = \begin{pmatrix} a \\ a \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where  $a$  is an arbitrary scalar. Only the nonzero solutions<sup>†</sup> count as eigenvectors, and so the eigenvectors for the eigenvalue  $\lambda_1 = 4$  must have  $a \neq 0$ , i.e., they are all nonzero scalar multiples of the basic eigenvector  $\mathbf{v}_1 = (1, 1)^T$ .

**Remark.** In general, if  $\mathbf{v}$  is an eigenvector of  $A$  for the eigenvalue  $\lambda$ , then so is every nonzero scalar multiple of  $\mathbf{v}$ . In practice, we distinguish only linearly independent eigenvectors. Thus, in this example, we shall say “ $\mathbf{v}_1 = (1, 1)^T$  is the eigenvector corresponding to the eigenvalue  $\lambda_1 = 4$ ”, when we really mean that the set of eigenvectors for  $\lambda_1 = 4$  consists of all nonzero scalar multiples of  $\mathbf{v}_1$ .

Similarly, for the second eigenvalue  $\lambda_2 = 2$ , the eigenvector equation is

$$(A - 2 \mathbf{I}) \mathbf{v} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The solution  $(-a, a)^T = a(-1, 1)^T$  is the set of scalar multiples of the eigenvector  $\mathbf{v}_2 = (-1, 1)^T$ . Therefore, the complete list of eigenvalues and eigenvectors (up to scalar multiple) for this particular matrix is

$$\lambda_1 = 4, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = 2, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

**Example 8.6.** Consider the  $3 \times 3$  matrix

$$A = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Using the formula (1.88) for a  $3 \times 3$  determinant, we compute the characteristic equation

$$\begin{aligned} 0 &= \det(A - \lambda \mathbf{I}) = \det \begin{pmatrix} -\lambda & -1 & -1 \\ 1 & 2 - \lambda & 1 \\ 1 & 1 & 2 - \lambda \end{pmatrix} \\ &= (-\lambda)(2 - \lambda)^2 + (-1) \cdot 1 \cdot 1 + (-1) \cdot 1 \cdot 1 \\ &\quad - 1 \cdot (2 - \lambda)(-1) - 1 \cdot 1 \cdot (-\lambda) - (2 - \lambda) \cdot 1 \cdot (-1) = -\lambda^3 + 4\lambda^2 - 5\lambda + 2. \end{aligned}$$

The resulting cubic polynomial can be factored:

$$-\lambda^3 + 4\lambda^2 - 5\lambda + 2 = -(\lambda - 1)^2(\lambda - 2) = 0.$$

---

<sup>†</sup> If, at this stage, you end up with a linear system with only the trivial zero solution, you've done something wrong! Either you don't have a correct eigenvalue — maybe you made a mistake setting up and/or solving the characteristic equation — or you've made an error solving the homogeneous eigenvector system. On the other hand, if you are working with a numerical approximation to the eigenvalue, then the resulting numerical homogeneous linear system will almost certainly not have a nonzero solution, and therefore a completely different approach must be taken to finding the corresponding eigenvector; see Sections 9.5 and 9.6.

Most  $3 \times 3$  matrices have three different eigenvalues, but this particular one has only two:  $\lambda_1 = 1$ , which is called a *double eigenvalue*, since it is a double root of the characteristic equation, along with a *simple eigenvalue*  $\lambda_2 = 2$ .

The eigenvector equation (8.13) for the double eigenvalue  $\lambda_1 = 1$  is

$$(A - I)\mathbf{v} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution to this homogeneous linear system

$$\mathbf{v} = \begin{pmatrix} -a - b \\ a \\ b \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

depends upon two free variables:  $y = a$  and  $z = b$ . Every nonzero solution forms a valid eigenvector for the eigenvalue  $\lambda_1 = 1$ , and so the general eigenvector is any non-zero linear combination of the two “basis eigenvectors”  $\mathbf{v}_1 = (-1, 1, 0)^T$ ,  $\hat{\mathbf{v}}_1 = (-1, 0, 1)^T$ .

On the other hand, the eigenvector equation for the simple eigenvalue  $\lambda_2 = 2$  is

$$(A - 2I)\mathbf{v} = \begin{pmatrix} -2 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution

$$\mathbf{v} = \begin{pmatrix} -a \\ a \\ a \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

consists of all scalar multiples of the eigenvector  $\mathbf{v}_2 = (-1, 1, 1)^T$ .

In summary, the eigenvalues and (basis) eigenvectors for this matrix are

$$\begin{aligned} \lambda_1 &= 1, & \mathbf{v}_1 &= \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, & \hat{\mathbf{v}}_1 &= \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \\ \lambda_2 &= 2, & \mathbf{v}_2 &= \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned} \tag{8.15}$$

This means that every eigenvector for the simple eigenvalue  $\lambda_2 = 2$  is a nonzero scalar multiple of  $\mathbf{v}_2$ , while every eigenvector for the double eigenvalue  $\lambda_1 = 1$  is a nontrivial linear combination  $a\mathbf{v}_1 + b\hat{\mathbf{v}}_1$  of the two linearly independent eigenvectors  $\mathbf{v}_1, \hat{\mathbf{v}}_1$ .

In general, given a real eigenvalue  $\lambda$ , the corresponding *eigenspace*  $V_\lambda \subset \mathbb{R}^n$  is the subspace spanned by all its eigenvectors. Equivalently, the eigenspace is the kernel

$$V_\lambda = \ker(A - \lambda I). \tag{8.16}$$

Thus,  $\lambda \in \mathbb{R}$  is an eigenvalue if and only if  $V_\lambda \neq \{\mathbf{0}\}$  is a nontrivial subspace, and then every nonzero element of  $V_\lambda$  is a corresponding eigenvector. The most economical way to indicate each eigenspace is by writing out a basis, as in (8.15), with  $\mathbf{v}_1, \hat{\mathbf{v}}_1$  giving a basis for the eigenspace  $V_1$ , while  $\mathbf{v}_2$  is a basis for the eigenspace  $V_2$ . In particular, 0 is an eigenvalue if and only if  $\ker A \neq \{\mathbf{0}\}$ , and hence  $A$  is singular.

**Proposition 8.7.** A matrix is singular if and only if it has a zero eigenvalue.

**Example 8.8.** The characteristic equation of the matrix  $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$  is

$$0 = \det(A - \lambda I) = -\lambda^3 + \lambda^2 + 5\lambda + 3 = -(\lambda + 1)^2(\lambda - 3).$$

Again, there is a double eigenvalue  $\lambda_1 = -1$  and a simple eigenvalue  $\lambda_2 = 3$ . However, in this case the matrix

$$A - \lambda_1 I = A + I = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \end{pmatrix}$$

has a one-dimensional kernel, spanned by  $\mathbf{v}_1 = (2, -1, -2)^T$ . Thus, even though  $\lambda_1$  is a double eigenvalue, it admits only a one-dimensional eigenspace. The list of eigenvalues and eigenvectors is, in a sense, incomplete:

$$\lambda_1 = -1, \quad \mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}, \quad \lambda_2 = 3, \quad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}.$$

**Example 8.9.** Finally, the matrix  $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -2 \\ 2 & 2 & -1 \end{pmatrix}$  has characteristic equation

$$0 = \det(A - \lambda I) = -\lambda^3 + \lambda^2 - 3\lambda - 5 = -(\lambda + 1)(\lambda^2 - 2\lambda + 5).$$

The linear factor yields the eigenvalue  $-1$ . The quadratic factor leads to two complex roots,  $1 + 2i$  and  $1 - 2i$ , which can be obtained via the quadratic formula. Hence  $A$  has one real and two complex eigenvalues:

$$\lambda_1 = -1, \quad \lambda_2 = 1 + 2i, \quad \lambda_3 = 1 - 2i.$$

On solving the associated linear system  $(A + I)\mathbf{v} = \mathbf{0}$ , the real eigenvalue  $\lambda_1 = -1$  is found to have corresponding eigenvector  $\mathbf{v}_1 = (-1, 1, 1)^T$ .

Complex eigenvalues are as important as real eigenvalues, and we need to be able to handle them too. To find the corresponding eigenvectors, which will also be complex, we need to solve the usual eigenvalue equation (8.13), which is now a complex homogeneous linear system. For example, the eigenvectors for  $\lambda_2 = 1 + 2i$  are found by solving

$$[A - (1 + 2i)I]\mathbf{v} = \begin{pmatrix} -2i & 2 & 0 \\ 0 & -2i & -2 \\ 2 & 2 & -2 - 2i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This linear system can be solved by Gaussian Elimination (with complex pivots). In this case, a simpler strategy is to work directly: the first equation  $-2ix + 2y = 0$  tells us that  $y = ix$ , while the second equation  $-2iy - 2z = 0$  says  $z = -iy = x$ . If we trust our calculations so far, we do not need to solve the final equation  $2x + 2y + (-2 - 2i)z = 0$ , since we know that the coefficient matrix is singular and hence this equation must be a consequence of the first two. (However, it does serve as a useful check on our work.) So, the general solution  $\mathbf{v} = (x, ix, x)^T$  is an arbitrary constant multiple of the complex eigenvector  $\mathbf{v}_2 = (1, i, 1)^T$ . The eigenvector equation for  $\lambda_3 = 1 - 2i$  is similarly solved for the third eigenvector  $\mathbf{v}_3 = (1, -i, 1)^T$ .

Summarizing, the matrix under consideration has one real and two complex eigenvalues, with three corresponding eigenvectors, each unique up to (complex) scalar multiple:

$$\begin{aligned}\lambda_1 &= -1, & \lambda_2 &= 1 + 2i, & \lambda_3 &= 1 - 2i, \\ \mathbf{v}_1 &= \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 1 \\ -i \\ 1 \end{pmatrix}.\end{aligned}$$

Note that the third complex eigenvalue is the complex conjugate of the second, and the eigenvectors are similarly related. This is indicative of a general fact for real matrices:

**Proposition 8.10.** If  $A$  is a real matrix with a complex eigenvalue  $\lambda = \mu + i\nu$  and corresponding complex eigenvector  $\mathbf{v} = \mathbf{x} + i\mathbf{y}$ , then the complex conjugate  $\bar{\lambda} = \mu - i\nu$  is also an eigenvalue with complex conjugate eigenvector  $\bar{\mathbf{v}} = \mathbf{x} - i\mathbf{y}$ .

*Proof:* First take complex conjugates of the eigenvalue equation (8.12):

$$\bar{A}\bar{\mathbf{v}} = \overline{A\mathbf{v}} = \overline{\lambda\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}.$$

Using the fact that a real matrix is unaffected by complex conjugation,  $\bar{A} = A$ , we conclude  $A\bar{\mathbf{v}} = \bar{\lambda}\bar{\mathbf{v}}$ , which is the equation for the eigenvalue  $\bar{\lambda}$  and eigenvector  $\bar{\mathbf{v}} \neq \mathbf{0}$ . *Q.E.D.*

As a consequence, when dealing with real matrices, we need to compute the eigenvectors for only *one* of each complex conjugate pair of eigenvalues. This observation effectively halves the amount of work in the unfortunate event that we are confronted with complex eigenvalues.

The *eigenspace* associated with a complex eigenvalue  $\lambda$  is the subspace  $V_\lambda \subset \mathbb{C}^n$  spanned by the corresponding (complex) eigenvectors. One might also consider complex eigenvectors associated with a real eigenvalue, but this doesn't add anything to the picture — they are merely complex linear combinations of the real eigenvectors. Thus, we need to introduce complex eigenvectors only when dealing with genuinely complex eigenvalues.

**Remark.** The reader may recall that we said that one should never use determinants in practical computations. So why have we reverted to using determinants to find eigenvalues? The truthful answer is that the practical computation of eigenvalues and eigenvectors *never* resorts to the characteristic equation! The method is fraught with numerical traps and inefficiencies when (a) computing the determinant leading to the characteristic equation, then (b) solving the resulting polynomial equation, which is itself a nontrivial numerical problem<sup>†</sup>, [8, 66], and, finally, (c) solving each of the resulting linear eigenvector systems. Even worse, if we know only an approximation  $\tilde{\lambda}$  to the true eigenvalue  $\lambda$ , the approximate eigenvector system  $(A - \tilde{\lambda}\mathbf{I})\mathbf{v} = \mathbf{0}$  will almost certainly have a nonsingular coefficient matrix, and hence admits only the trivial solution  $\mathbf{v} = \mathbf{0}$  — which does not even qualify as an eigenvector!

Nevertheless, the characteristic equation does give us important theoretical insight into the structure of the eigenvalues of a matrix, and can be used when dealing with very small matrices, e.g.,  $2 \times 2$  and  $3 \times 3$ , presuming exact arithmetic is employed. Numerical

---

<sup>†</sup> In fact, one effective numerical strategy for finding the roots of a polynomial is to turn the procedure on its head, and calculate the eigenvalues of a matrix whose characteristic equation is the polynomial in question! See [66] for details.

algorithms for computing eigenvalues and eigenvectors are based on completely different ideas, and will be deferred until Sections 9.5 and 9.6.

## Exercises

8.2.1. Find the eigenvalues and eigenvectors of the following matrices:

- (a)  $\begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & -\frac{2}{3} \\ \frac{1}{2} & -\frac{1}{6} \end{pmatrix}$ , (c)  $\begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$ ,  
 (e)  $\begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$ , (f)  $\begin{pmatrix} -1 & -1 & 4 \\ 1 & 3 & -2 \\ 1 & 1 & -1 \end{pmatrix}$ , (g)  $\begin{pmatrix} 1 & -3 & 11 \\ 2 & -6 & 16 \\ 1 & -3 & 7 \end{pmatrix}$ , (h)  $\begin{pmatrix} 2 & -1 & -1 \\ -2 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ ,  
 (i)  $\begin{pmatrix} -4 & -4 & 2 \\ 3 & 4 & -1 \\ -3 & -2 & 3 \end{pmatrix}$ , (j)  $\begin{pmatrix} 3 & 4 & 0 & 0 \\ 4 & 3 & 0 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 4 & 5 \end{pmatrix}$ , (k)  $\begin{pmatrix} 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ -1 & 1 & 2 & 0 \\ 1 & -1 & 1 & 1 \end{pmatrix}$ .

8.2.2. (a) Find the eigenvalues of the rotation matrix  $R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$ . For what values of  $\theta$  are the eigenvalues real?

(b) Explain why your answer gives an immediate solution to Exercise 1.5.7c.

8.2.3. Answer Exercise 8.2.2a for the reflection matrix  $F_\theta = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}$ .

8.2.4. Write down (a) a  $2 \times 2$  matrix that has 0 as one of its eigenvalues and  $(1, 2)^T$  as a corresponding eigenvector; (b) a  $3 \times 3$  matrix that has  $(1, 2, 3)^T$  as an eigenvector for the eigenvalue  $-1$ .

8.2.5. (a) Write out the characteristic equation for the matrix  $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \alpha & \beta & \gamma \end{pmatrix}$ .

(b) Show that, given any 3 numbers  $a$ ,  $b$ , and  $c$ , there is a  $3 \times 3$  matrix with characteristic equation  $-\lambda^3 + a\lambda^2 + b\lambda + c = 0$ .

8.2.6. Find the eigenvalues and eigenvectors of the cross product matrix  $A = \begin{pmatrix} 0 & c & -b \\ -c & 0 & a \\ b & -a & 0 \end{pmatrix}$ .

8.2.7. Find all eigenvalues and eigenvectors of the following complex matrices:

- (a)  $\begin{pmatrix} i & 1 \\ 0 & -1+i \end{pmatrix}$ , (b)  $\begin{pmatrix} 2 & i \\ -i & -2 \end{pmatrix}$ , (c)  $\begin{pmatrix} i-2 & i+1 \\ i+2 & i-1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1+i & -1-i & 1-i \\ 2 & -2-i & 2-2i \\ -1 & 1+i & -1+2i \end{pmatrix}$ .

8.2.8. Find all eigenvalues and eigenvectors of

- (a) the  $n \times n$  zero matrix  $O$ ; (b) the  $n \times n$  identity matrix  $I$ .

8.2.9. Find the eigenvalues and eigenvectors of an  $n \times n$  matrix with every entry equal to 1.

*Hint:* Try with  $n = 2, 3$ , and then generalize.

◇ 8.2.10. Let  $A$  be a given square matrix. (a) Explain in detail why every nonzero scalar multiple of an eigenvector of  $A$  is also an eigenvector. (b) Show that every nonzero linear combination of two eigenvectors  $\mathbf{v}, \mathbf{w}$  corresponding to the *same* eigenvalue is also an eigenvector. (c) Prove that a linear combination  $c\mathbf{v} + d\mathbf{w}$ , with  $c, d \neq 0$ , of two eigenvectors corresponding to *different* eigenvalues is never an eigenvector.

◇ 8.2.11. Let  $\lambda$  be a real eigenvalue of the real  $n \times n$  matrix  $A$ , and  $\mathbf{v}_1, \dots, \mathbf{v}_k$  a basis for the associated eigenspace  $V_\lambda$ . Suppose  $\mathbf{w} \in \mathbb{C}^n$  is a complex eigenvector, so  $A\mathbf{w} = \lambda\mathbf{w}$ . Prove that  $\mathbf{w} = c_1\mathbf{v}_1 + \dots + c_k\mathbf{v}_k$  is a complex linear combination of the real eigenspace basis.

*Hint:* Look at the real and imaginary parts of the eigenvector equation.

8.2.12. *True or false:* If  $\mathbf{v}$  is a real eigenvector of a real matrix  $A$ , then a nonzero complex multiple  $\mathbf{w} = c\mathbf{v}$  for  $c \in \mathbb{C}$  is a complex eigenvector of  $A$ .

◇ 8.2.13. Define the *shift map*  $S: \mathbb{C}^n \rightarrow \mathbb{C}^n$  by  $S(v_1, v_2, \dots, v_{n-1}, v_n)^T = (v_2, v_3, \dots, v_n, v_1)^T$ .

(a) Prove that  $S$  is a linear map, and write down its matrix representation  $A$ .

(b) Prove that  $A$  is an orthogonal matrix. (c) Prove that the sampled exponential vectors  $\omega_0, \dots, \omega_{n-1}$  defined in (5.102) form an eigenvector basis of  $A$ . What are the eigenvalues?

## Basic Properties of Eigenvalues

If  $A$  is an  $n \times n$  matrix, then its *characteristic polynomial* is defined to be

$$p_A(\lambda) = \det(A - \lambda I) = c_n \lambda^n + c_{n-1} \lambda^{n-1} + \dots + c_1 \lambda + c_0. \quad (8.17)$$

The fact that  $p_A(\lambda)$  is a polynomial of degree  $n$  is a consequence of the general determinantal formula (1.87). Indeed, every term is prescribed by a permutation  $\pi$  of the rows of the matrix, and equals plus or minus a product of  $n$  distinct matrix entries including one from each row and one from each column. The term corresponding to the identity permutation is obtained by multiplying the diagonal entries together, which, in this case, is

$$(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda) = (-1)^n \lambda^n + (-1)^{n-1} (a_{11} + a_{22} + \dots + a_{nn}) \lambda^{n-1} + \dots \quad (8.18)$$

All of the other terms have at most  $n - 2$  diagonal factors  $a_{ii} - \lambda$ , and so are polynomials of degree  $\leq n - 2$  in  $\lambda$ . Thus, (8.18) is the only summand containing the monomials  $\lambda^n$  and  $\lambda^{n-1}$ , and so their respective coefficients are

$$c_n = (-1)^n, \quad c_{n-1} = (-1)^{n-1} (a_{11} + a_{22} + \dots + a_{nn}) = (-1)^{n-1} \operatorname{tr} A, \quad (8.19)$$

where  $\operatorname{tr} A$ , the sum of its diagonal entries, is called the *trace* of the matrix  $A$ . The other coefficients  $c_{n-2}, \dots, c_1, c_0$  in (8.17) are more complicated combinations of the entries of  $A$ . However, setting  $\lambda = 0$  implies

$$p_A(0) = c_0 = \det A, \quad (8.20)$$

and hence the constant term in the characteristic polynomial equals the determinant of the matrix. In particular, if  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is a  $2 \times 2$  matrix, its characteristic polynomial has the explicit form

$$\begin{aligned} p_A(\lambda) &= \det(A - \lambda I) = \det \begin{pmatrix} a - \lambda & b \\ c & d - \lambda \end{pmatrix} \\ &= \lambda^2 - (a + d)\lambda + (ad - bc) = \lambda^2 - (\operatorname{tr} A)\lambda + (\det A). \end{aligned} \quad (8.21)$$

According to the Fundamental Theorem of Algebra, [26], every (complex) polynomial of degree  $n \geq 1$  can be completely factored, and so we can write the characteristic polynomial (8.17) in factored form:

$$p_A(\lambda) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n). \quad (8.22)$$

The complex numbers  $\lambda_1, \dots, \lambda_n$ , some of which may be repeated, are the *roots* of the characteristic equation  $p_A(\lambda) = 0$ , and hence the eigenvalues of the matrix  $A$ . Therefore, we immediately conclude:

**Theorem 8.11.** An  $n \times n$  matrix possesses at least one and at most  $n$  distinct complex eigenvalues.

Most  $n \times n$  matrices — meaning those for which the characteristic polynomial factors into  $n$  *distinct* factors — have *exactly*  $n$  complex eigenvalues. More generally, an eigenvalue  $\lambda_j$  is said to have *multiplicity*  $k$  if the factor  $(\lambda - \lambda_j)$  appears exactly  $k$  times in the factorization (8.22) of the characteristic polynomial. An eigenvalue is *simple* if it has multiplicity 1, *double* if it has multiplicity 2, and so on. In particular,  $A$  has  $n$  distinct eigenvalues if and only if all its eigenvalues are simple. In all cases, when the repeated eigenvalues are counted in accordance with their multiplicity, every  $n \times n$  matrix has a total of  $n$  complex eigenvalues.

An example of a matrix with just one eigenvalue, of multiplicity  $n$ , is the  $n \times n$  identity matrix  $I$ , whose only eigenvalue is  $\lambda = 1$ . In this case, *every* nonzero vector in  $\mathbb{R}^n$  is an eigenvector of the identity matrix (why?), and so the eigenspace  $V_1$  is all of  $\mathbb{R}^n$ . At the other extreme, the  $n \times n$  “bidiagonal” *Jordan block matrix*<sup>†</sup>

$$J_{a,n} = \begin{pmatrix} a & 1 & & & \\ & a & 1 & & \\ & & a & 1 & \\ & & & \ddots & \ddots \\ & & & & a & 1 \\ & & & & & a \end{pmatrix} \quad (8.23)$$

also has only one eigenvalue,  $\lambda = a$ , again of multiplicity  $n$ . But in this case,  $J_{a,n}$  has only one eigenvector (up to scalar multiple), which is the first standard basis vector  $\mathbf{e}_1$ , and so its eigenspace is one-dimensional.

**Remark.** If  $\lambda$  is a complex eigenvalue of multiplicity  $k$  for the real matrix  $A$ , then its complex conjugate  $\bar{\lambda}$  also has multiplicity  $k$ . This is because complex conjugate roots of a real polynomial necessarily appear with identical multiplicities.

**Remark.** If  $n \leq 4$ , then one can, in fact, write down an explicit formula for the solution to a polynomial equation of degree  $n$ , and hence explicit (but rather complicated and not particularly helpful) formulas for the eigenvalues of general  $2 \times 2$ ,  $3 \times 3$ , and  $4 \times 4$  matrices. As soon as  $n \geq 5$ , there is no explicit formula (at least in terms of radicals), and so one must usually resort to numerical approximations. This remarkable and deep algebraic result was proved in the early nineteenth century by the young Norwegian mathematician Niels Henrik Abel, [26].

**Proposition 8.12.** A square matrix  $A$  and its transpose  $A^T$  have the same characteristic equation, and hence the same eigenvalues with the same multiplicities.

*Proof:* This follows immediately from Proposition 1.56, that the determinant of a matrix and its transpose are identical. Thus,

$$p_A(\lambda) = \det(A - \lambda I) = \det(A - \lambda I)^T = \det(A^T - \lambda I) = p_{A^T}(\lambda). \quad Q.E.D.$$

**Remark.** While  $A^T$  has the same eigenvalues as  $A$ , its eigenvectors are, in general, different. An eigenvector  $\mathbf{v}$  of  $A^T$ , satisfying  $A^T \mathbf{v} = \lambda \mathbf{v}$ , is sometimes referred to as a *left eigenvector* of  $A$ , since it satisfies  $\mathbf{v}^T A = \lambda \mathbf{v}^T$ . A more apt, albeit rather non-conventional, name for  $\mathbf{v}$  that conforms with our nomenclature conventions would be *co-eigenvector*.

---

<sup>†</sup> All non-displayed entries are zero.

If we explicitly multiply out the factored product (8.22) and equate the result to the characteristic polynomial (8.17), we find that its coefficients  $c_0, c_1, \dots, c_{n-1}$  can be written as certain polynomials of the roots, known as the *elementary symmetric polynomials*. The first and last are of particular importance:

$$c_0 = \lambda_1 \lambda_2 \cdots \lambda_n, \quad c_{n-1} = (-1)^{n-1} (\lambda_1 + \lambda_2 + \cdots + \lambda_n). \quad (8.24)$$

Comparison with our previous formulas (8.19, 20) for the coefficients  $c_0$  and  $c_{n-1}$  leads to the following useful result.

**Proposition 8.13.** The sum of the eigenvalues of a square matrix  $A$  equals its trace:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \text{tr } A = a_{11} + a_{22} + \cdots + a_{nn}. \quad (8.25)$$

The product of the eigenvalues equals its determinant:

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det A. \quad (8.26)$$

Keep in mind that, in evaluating (8.25, 26), one must add or multiply repeated eigenvalues according to their multiplicity.

**Example 8.14.** The matrix  $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$  considered in Example 8.8 has trace and determinant

$$\text{tr } A = 1, \quad \det A = 3,$$

which fix, respectively, the coefficient of  $\lambda^2$  and the constant term in its characteristic equation. This matrix has two distinct eigenvalues:  $-1$ , which is a double eigenvalue, and  $3$ , which is simple. For this particular matrix, (8.25, 26) become

$$1 = \text{tr } A = (-1) + (-1) + 3, \quad 3 = \det A = (-1)(-1)3.$$

Note that the double eigenvalue contributes twice to both the sum and the product.

## Exercises

8.2.14. (a) Compute the eigenvalues and corresponding eigenvectors of  $A = \begin{pmatrix} 1 & 4 & 4 \\ 3 & -1 & 0 \\ 0 & 2 & 3 \end{pmatrix}$ .

(b) Compute the trace of  $A$  and check that it equals the sum of the eigenvalues. (c) Find the determinant of  $A$  and check that it is equal to the product of the eigenvalues.

8.2.15. Verify the trace and determinant formulas (8.25, 26) for the matrices in Exercise 8.2.1.

8.2.16. (a) Find the explicit formula for the characteristic polynomial

$\det(A - \lambda I) = -\lambda^3 + a\lambda^2 - b\lambda + c$  of a general  $3 \times 3$  matrix. Verify that  $a = \text{tr } A$ ,  $c = \det A$ . What is the formula for  $b$ ? (b) Prove that if  $A$  has eigenvalues  $\lambda_1, \lambda_2, \lambda_3$ , then  $a = \text{tr } A = \lambda_1 + \lambda_2 + \lambda_3$ ,  $b = \lambda_1 \lambda_2 + \lambda_1 \lambda_3 + \lambda_2 \lambda_3$ ,  $c = \det A = \lambda_1 \lambda_2 \lambda_3$ .

8.2.17. Prove that the eigenvalues of an upper triangular (or lower triangular) matrix are its diagonal entries.

◇ 8.2.18. Let  $J_{a,n}$  be the  $n \times n$  Jordan block matrix (8.23). Prove that its only eigenvalue is  $\lambda = a$  and the only eigenvectors are the nonzero scalar multiples of the standard basis vector  $\mathbf{e}_1$ .

- ◇ 8.2.19. Suppose that  $\lambda$  is an eigenvalue of  $A$ . (a) Prove that  $c\lambda$  is an eigenvalue of the scalar multiple  $cA$ . (b) Prove that  $\lambda + d$  is an eigenvalue of  $A + dI$ . (c) More generally,  $c\lambda + d$  is an eigenvalue of  $B = cA + dI$  for scalars  $c, d$ .
- 8.2.20. Show that if  $\lambda$  is an eigenvalue of  $A$ , then  $\lambda^2$  is an eigenvalue of  $A^2$ .
- 8.2.21. *True or false:* (a) If  $\lambda$  is an eigenvalue of both  $A$  and  $B$ , then it is an eigenvalue of the sum  $A + B$ . (b) If  $\mathbf{v}$  is an eigenvector of both  $A$  and  $B$ , then it is an eigenvector of  $A + B$ .
- 8.2.22. *True or false:* If  $\lambda$  is an eigenvalue of  $A$  and  $\mu$  is an eigenvalue of  $B$ , then  $\lambda\mu$  is an eigenvalue of the matrix product  $C = AB$ .
- ◇ 8.2.23. Let  $A$  and  $B$  be  $n \times n$  matrices. Prove that the matrix products  $AB$  and  $BA$  have the same eigenvalues. *Hint:* How should the eigenvectors be related?
- ◇ 8.2.24. (a) Prove that if  $\lambda \neq 0$  is a nonzero eigenvalue of the nonsingular matrix  $A$ , then  $1/\lambda$  is an eigenvalue of  $A^{-1}$ . (b) What happens if  $A$  has 0 as an eigenvalue?
- ◇ 8.2.25. (a) Prove that if  $|\det A| > 1$ , then  $A$  has at least one eigenvalue with  $|\lambda| > 1$ . (b) If  $|\det A| < 1$ , are all eigenvalues  $|\lambda| < 1$ ? Prove or find a counterexample.
- 8.2.26. Prove that  $A$  is a singular matrix if and only if 0 is an eigenvalue.
- 8.2.27. Prove that *every* nonzero vector  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$  is an eigenvector of  $A$  if and only if  $A$  is a scalar multiple of the identity matrix.
- 8.2.28. How many unit (norm 1) eigenvectors correspond to a given eigenvalue of a matrix?
- 8.2.29. *True or false:* (a) Performing an elementary row operation of type #1 does not change the eigenvalues of a matrix. (b) Interchanging two rows of a matrix changes the sign of its eigenvalues. (c) Multiplying one row of a matrix by a scalar multiplies one of its eigenvalues by the same scalar.
- 8.2.30. (a) *True or false:* If  $\lambda_1, \mathbf{v}_1$  and  $\lambda_2, \mathbf{v}_2$  solve the eigenvalue equation (8.12) for a given matrix  $A$ , so does  $\lambda_1 + \lambda_2, \mathbf{v}_1 + \mathbf{v}_2$ . (b) Explain what this has to do with linearity.
- 8.2.31. As in (4.35), an elementary reflection matrix has the form  $Q = I - 2\mathbf{u}\mathbf{u}^T$ , where  $\mathbf{u} \in \mathbb{R}^n$  is a unit vector. (a) Find the eigenvalues and eigenvectors of the elementary reflection matrices for the unit vectors (i)  $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ , (ii)  $\begin{pmatrix} \frac{3}{5} \\ \frac{4}{5} \end{pmatrix}$ , (iii)  $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ , (iv)  $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$ . (b) What are the eigenvalues and eigenvectors of a general elementary reflection matrix?
- ◇ 8.2.32. Let  $A$  and  $B$  be similar matrices, so  $B = S^{-1}AS$  for some nonsingular matrix  $S$ . (a) Prove that  $A$  and  $B$  have the same characteristic polynomial:  $p_B(\lambda) = p_A(\lambda)$ . (b) Explain why similar matrices have the same eigenvalues. (c) Do they have the same eigenvectors? If not, how are their eigenvectors related? (d) Prove that the converse to part (c) is false by showing that  $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$  and  $\begin{pmatrix} 1 & 1 \\ -1 & 3 \end{pmatrix}$  have the same eigenvalues, but are *not* similar.
- 8.2.33. Let  $A$  be a nonsingular  $n \times n$  matrix with characteristic polynomial  $p_A(\lambda)$ . (a) Explain how to construct the characteristic polynomial  $p_{A^{-1}}(\lambda)$  of its inverse directly from  $p_A(\lambda)$ . (b) Check your result when  $A =$  (i)  $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 1 & 4 & 4 \\ -2 & -1 & 0 \\ 0 & 2 & 3 \end{pmatrix}$ .
- ◇ 8.2.34. Prove that the only eigenvalue of a nilpotent matrix, cf. Exercise 1.3.12, is 0. (The converse is also true; see Exercise 8.6.20.)

8.2.35. Given an idempotent matrix, so that  $P = P^2$ , find all its eigenvalues and eigenvectors.

8.2.36. (a) Prove that every real  $3 \times 3$  matrix has at least one real eigenvalue.

(b) Find a real  $4 \times 4$  matrix with no real eigenvalues.

(c) Can you find a real  $5 \times 5$  matrix with no real eigenvalues?

8.2.37. (a) Show that if  $A$  is a matrix such that  $A^4 = I$ , then the only possible eigenvalues of  $A$  are  $1, -1, i$ , and  $-i$ .

(b) Give an example of a real matrix that has all four numbers as eigenvalues.

8.2.38. (a) Prove that if  $\lambda$  is an eigenvalue of  $A$ , then  $\lambda^n$  is an eigenvalue of  $A^n$ . (b) State and prove a converse.

8.2.39. *True or false:* All the eigenvalues of an  $n \times n$  permutation matrix are real.

8.2.40. (a) Show that if all the row sums of  $A$  are equal to 1, then  $A$  has 1 as an eigenvalue.

(b) Suppose all the column sums of  $A$  are equal to 1. Does the same result hold?

◇ 8.2.41. Prove that if  $\mathbf{v}$  is an eigenvector of  $A$  with eigenvalue  $\lambda$  and  $\mathbf{w}$  is an eigenvector of  $A^T$  with a different eigenvalue  $\mu \neq \lambda$ , then  $\mathbf{v}$  and  $\mathbf{w}$  are orthogonal vectors with respect to the

dot product. Illustrate this result when (i)  $A = \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix}$ , (ii)  $A = \begin{pmatrix} 5 & -4 & 2 \\ 5 & -4 & 1 \\ -2 & 2 & -3 \end{pmatrix}$ .

8.2.42. Let  $Q$  be an orthogonal matrix. (a) Prove that if  $\lambda$  is an eigenvalue, then so is  $1/\lambda$ .

(b) Prove that all its eigenvalues are complex numbers of modulus  $|\lambda| = 1$ . In particular, the only possible real eigenvalues of an orthogonal matrix are  $\pm 1$ . (c) Suppose  $\mathbf{v} = \mathbf{x} + i\mathbf{y}$  is a complex eigenvector corresponding to a non-real eigenvalue. Prove that its real and imaginary parts are orthogonal vectors having the same Euclidean norm.

◇ 8.2.43. (a) Prove that every  $3 \times 3$  proper orthogonal matrix has  $+1$  as an eigenvalue.

(b) *True or false:* An improper  $3 \times 3$  orthogonal matrix has  $-1$  as an eigenvalue.

◇ 8.2.44. (a) Show that the linear transformation defined by a  $3 \times 3$  proper orthogonal matrix corresponds to rotating through an angle around a line through the origin in  $\mathbb{R}^3$  — the axis of the rotation. *Hint:* Use Exercise 8.2.43(a).

(b) Find the axis and angle of rotation of the orthogonal matrix  $\begin{pmatrix} \frac{3}{5} & 0 & \frac{4}{5} \\ -\frac{4}{13} & \frac{12}{13} & \frac{3}{13} \\ -\frac{48}{65} & -\frac{5}{13} & \frac{36}{65} \end{pmatrix}$ .

◇ 8.2.45. Prove that every *proper* affine isometry  $F(\mathbf{x}) = Q\mathbf{x} + \mathbf{b}$  of  $\mathbb{R}^3$ , where  $\det Q = +1$ , is one of the following: (a) a *translation*  $\mathbf{x} + \mathbf{b}$ , (b) a *rotation* centered at some point of  $\mathbb{R}^3$ , or (c) a *screw motion* consisting of a rotation around an axis followed by a translation in the direction of the axis. *Hint:* Use Exercise 8.2.44.

8.2.46. Suppose  $Q$  is an orthogonal matrix. (a) Prove that  $K = 2I - Q - Q^T$  is a positive semi-definite matrix. (b) Under what conditions is  $K > 0$ ?

◇ 8.2.47. Let  $M_n$  be the  $n \times n$  tridiagonal matrix whose diagonal entries are all equal to 0 and whose sub- and super-diagonal entries all equal 1. (a) Find the eigenvalues and eigenvectors of  $M_2$  and  $M_3$  directly. (b) Prove that the eigenvalues and eigenvectors of  $M_n$  are explicitly given by

$$\lambda_k = 2 \cos \frac{k\pi}{n+1}, \quad \mathbf{v}_k = \left( \sin \frac{k\pi}{n+1}, \quad \sin \frac{2k\pi}{n+1}, \quad \dots \quad \sin \frac{nk\pi}{n+1} \right)^T, \quad k = 1, \dots, n.$$

How do you know that there are no other eigenvalues?

◇ 8.2.48. Let  $a, b \in \mathbb{R}$ . Determine the eigenvalues and eigenvectors of the  $n \times n$  tridiagonal matrix with all diagonal entries equal to  $a$  and all sub- and super-diagonal entries equal to  $b$ . *Hint:* See Exercises 8.2.19 and 8.2.47.

- ◇ 8.2.49. Find a formula for the eigenvalues of the tricirculant  $n \times n$  matrix  $Z_n$  that has 1's on the sub- and super-diagonals as well as its  $(1, n)$  and  $(n, 1)$  entries, while all other entries are 0. *Hint:* Use Exercise 8.2.47 as a guide.
- ◇ 8.2.50. Let  $A$  be an  $n \times n$  matrix with eigenvalues  $\lambda_1, \dots, \lambda_k$ , and  $B$  an  $m \times m$  matrix with eigenvalues  $\mu_1, \dots, \mu_l$ . Show that the  $(m+n) \times (m+n)$  block diagonal matrix  $D = \begin{pmatrix} A & O \\ O & B \end{pmatrix}$  has eigenvalues  $\lambda_1, \dots, \lambda_k, \mu_1, \dots, \mu_l$  and no others. How are the eigenvectors related?
- ◇ 8.2.51. *Deflation:* Suppose  $A$  has eigenvalue  $\lambda$  and corresponding eigenvector  $\mathbf{v}$ . (a) Let  $\mathbf{b}$  be any vector. Prove that the matrix  $B = A - \mathbf{v}\mathbf{b}^T$  also has  $\mathbf{v}$  as an eigenvector, now with eigenvalue  $\lambda - \beta$ , where  $\beta = \mathbf{v} \cdot \mathbf{b}$ . (b) Prove that if  $\mu \neq \lambda - \beta$  is any other eigenvalue of  $A$ , then it is also an eigenvalue of  $B$ . *Hint:* Look for an eigenvector of the form  $\mathbf{w} + c\mathbf{v}$ , where  $\mathbf{w}$  is an eigenvector of  $A$ . (c) Given a nonsingular matrix  $A$  with eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_n$  and  $\lambda_1 \neq \lambda_j$  for all  $j \geq 2$ , explain how to construct a *deflated* matrix  $B$  whose eigenvalues are  $0, \lambda_2, \dots, \lambda_n$ . (d) Try out your method on the matrices  $\begin{pmatrix} 3 & 3 \\ 1 & 5 \end{pmatrix}$  and  $\begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$ .
- ◇ 8.2.52. Let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  be a  $2 \times 2$  matrix. (a) Prove that  $A$  satisfies its own characteristic equation, meaning  $p_A(A) = A^2 - (\text{tr } A)A + (\det A)\mathbf{I} = \mathbf{O}$ . **Remark.** This result is a special case of the *Cayley–Hamilton Theorem*, to be developed in Exercise 8.6.22.
- (b) Prove the inverse formula  $A^{-1} = \frac{(\text{tr } A)\mathbf{I} - A}{\det A}$  when  $\det A \neq 0$ .
- (c) Check the Cayley–Hamilton and inverse formulas when  $A = \begin{pmatrix} 2 & 1 \\ -3 & 2 \end{pmatrix}$ .

## The Gershgorin Circle Theorem

In general, precisely computing the eigenvalues of a matrix is not easy, and, in most cases, must be done through a numerical eigenvalue procedure; see Sections 9.5 and 9.6. In certain applications, though, we may not require exact numerical values, but only their approximate locations. The *Gershgorin Circle Theorem*, due to the early-twentieth-century Russian mathematician Semyon Gershgorin, serves to restrict the eigenvalues to a certain well-defined region in the complex plane.

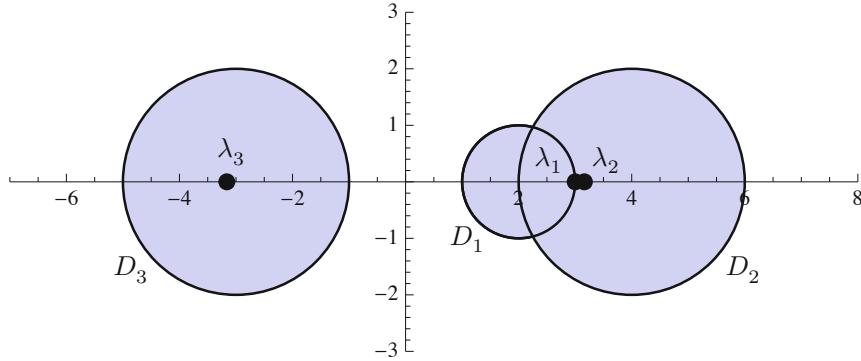
**Definition 8.15.** Let  $A$  be an  $n \times n$  matrix, either real or complex. For each  $1 \leq i \leq n$ , define the  $i^{\text{th}}$  *Gershgorin disk*

$$D_i = \{ |z - a_{ii}| \leq r_i \mid z \in \mathbb{C} \}, \quad \text{where} \quad r_i = \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|. \quad (8.27)$$

The *Gershgorin domain*  $D_A = \bigcup_{i=1}^n D_i \subset \mathbb{C}$  is the union of the Gershgorin disks.

Thus, the  $i^{\text{th}}$  Gershgorin disk  $D_i$  is centered at the  $i^{\text{th}}$  diagonal entry  $a_{ii}$  of  $A$ , and has radius  $r_i$  equal to the sum of the absolute values of the off-diagonal entries that are in its  $i^{\text{th}}$  row. We can now state the Gershgorin Circle Theorem.

**Theorem 8.16.** All real and complex eigenvalues of the matrix  $A$  lie in its Gershgorin domain  $D_A \subset \mathbb{C}$ .



**Figure 8.2.** Gershgorin Disks and Eigenvalues.

**Example 8.17.** The matrix  $A = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 4 & -1 \\ -1 & -1 & -3 \end{pmatrix}$  has Gershgorin disks

$$D_1 = \{ |z - 2| \leq 1 \}, \quad D_2 = \{ |z - 4| \leq 2 \}, \quad D_3 = \{ |z + 3| \leq 2 \},$$

which are plotted in Figure 8.2. The eigenvalues of  $A$  are

$$\lambda_1 = 3, \quad \lambda_2 = \sqrt{10} = 3.1622\ldots, \quad \lambda_3 = -\sqrt{10} = -3.1622\ldots.$$

Observe that  $\lambda_1$  belongs to both  $D_1$  and  $D_2$ , while  $\lambda_2$  lies in  $D_2$ , and  $\lambda_3$  is in  $D_3$ . We thus confirm that all three eigenvalues are in the Gershgorin domain  $D_A = D_1 \cup D_2 \cup D_3$ .

*Proof of Theorem 8.16:* Let  $\mathbf{v}$  be an eigenvector of  $A$  with eigenvalue  $\lambda$ . Let  $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|_\infty$  be the corresponding unit eigenvector with respect to the  $\infty$  norm, so

$$\|\mathbf{u}\|_\infty = \max\{ |u_1|, \dots, |u_n| \} = 1.$$

Let  $u_i$  be an entry of  $\mathbf{u}$  that achieves the maximum:  $|u_i| = 1$ . Writing out the  $i^{\text{th}}$  component of the eigenvalue equation  $A\mathbf{u} = \lambda\mathbf{u}$ , we obtain

$$\sum_{j=1}^n a_{ij} u_j = \lambda u_i, \quad \text{which we rewrite as} \quad \sum_{j \neq i} a_{ij} u_j = (\lambda - a_{ii}) u_i.$$

Therefore, since all  $|u_j| \leq 1$ , while  $|u_i| = 1$ ,

$$|\lambda - a_{ii}| = |\lambda - a_{ii}| |u_i| = |(\lambda - a_{ii}) u_i| = \left| \sum_{j \neq i} a_{ij} u_j \right| \leq \sum_{j \neq i} |a_{ij}| |u_j| \leq \sum_{j \neq i} |a_{ij}| = r_i.$$

This immediately implies that  $\lambda \in D_i \subset D_A$  belongs to the  $i^{\text{th}}$  Gershgorin disk. *Q.E.D.*

According to Proposition 8.7, a matrix  $A$  is singular if and only if it admits zero as an eigenvalue. Thus, if its Gershgorin domain does not contain 0, it cannot be an eigenvalue, and hence  $A$  is necessarily invertible. The condition  $0 \notin D_A$  requires that the matrix have large diagonal entries, as quantified by the following definition.

**Definition 8.18.** A square matrix  $A$  is called *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|, \quad \text{for all } i = 1, \dots, n. \quad (8.28)$$

In other words, strict diagonal dominance requires each diagonal entry to be larger, in absolute value, than the sum of the absolute values of *all* the other entries in its row. For example, the matrix  $\begin{pmatrix} 3 & -1 & 1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}$  is strictly diagonally dominant since

$$|3| > |-1| + |1|, \quad |-4| > |1| + |2|, \quad |5| > |-2| + |-1|.$$

Diagonally dominant matrices appear frequently in numerical solution methods for both ordinary and partial differential equations.

**Theorem 8.19.** A strictly diagonally dominant matrix is nonsingular.

*Proof:* The diagonal dominance inequalities (8.28) imply that the radius of the  $i^{\text{th}}$  Gershgorin disk is strictly less than the modulus of its center:  $r_i < |a_{ii}|$ . This implies that the disk cannot contain 0; indeed, if  $z \in D_i$ , then, by the triangle inequality,

$$r_i > |z - a_{ii}| \geq |a_{ii}| - |z| > r_i - |z|, \quad \text{and hence} \quad |z| > 0.$$

Thus, 0 does not lie in the Gershgorin domain  $D_A$ , and so cannot be an eigenvalue. *Q.E.D.*

**Warning.** The converse to this result is obviously not true; there are plenty of nonsingular matrices that are not strictly diagonally dominant.

## Exercises

8.2.53. For each of the following matrices,

- (i) find all Gershgorin disks; (ii) plot the Gershgorin domain in the complex plane; (iii) compute the eigenvalues and confirm the truth of the Circle Theorem 8.16:

$$(a) \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & -\frac{2}{3} \\ \frac{1}{2} & -\frac{1}{6} \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix},$$

$$(e) \begin{pmatrix} -1 & 1 & 1 \\ 2 & 2 & -1 \\ 0 & 3 & -4 \end{pmatrix}, \quad (f) \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{6} & 0 \end{pmatrix}, \quad (g) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{pmatrix}, \quad (h) \begin{pmatrix} 3 & 2 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 2 & 1 \end{pmatrix}.$$

8.2.54. *True or false:* The Gershgorin domain of the transpose of a matrix  $A^T$  is the same as the Gershgorin domain of the matrix  $A$ , that is,  $D_{A^T} = D_A$ .

◇ 8.2.55.(i) Explain why the eigenvalues of  $A$  must lie in its *refined Gershgorin domain*  $D_A^* = D_{A^T} \cap D_A$ . (ii) Find the refined Gershgorin domains for each of the matrices in Exercise 8.2.53 and confirm the result in part (i).

8.2.56. *True or false:* (a) A positive definite matrix is strictly diagonally dominant. (b) A strictly diagonally dominant matrix is positive definite.

◇ 8.2.57. Prove that if  $K$  is symmetric, strictly diagonally dominant, and each diagonal entry is positive, then  $K$  is positive definite.

8.2.58. (a) Write down an invertible matrix  $A$  whose Gershgorin domain contains 0. (b) Can you find an example that is also strictly diagonally dominant?

### 8.3 Eigenvector Bases

Most of the vector space bases that play a distinguished role in applications are assembled from the eigenvectors of a particular matrix. In this section, we show that the eigenvectors of a “complete” matrix automatically form a basis for  $\mathbb{R}^n$ , or, in the complex case,  $\mathbb{C}^n$ . In the following subsection, we use the eigenvector basis to rewrite the linear transformation determined by the matrix in a simple diagonal form, hence the alternative more common term “diagonalizable” for such matrices. The most important cases — symmetric and positive definite matrices — will be treated in the following section.

The first task is to show that eigenvectors corresponding to distinct eigenvalues are automatically linearly independent.

**Lemma 8.20.** If  $\lambda_1, \dots, \lambda_k$  are *distinct* eigenvalues of a matrix  $A$ , so  $\lambda_i \neq \lambda_j$  when  $i \neq j$ , then the corresponding eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_k$  are linearly independent.

*Proof:* The result is proved by induction on the number of eigenvalues. The case  $k = 1$  is immediate, since an eigenvector cannot be zero. Assume that we know that the result is valid for  $k - 1$  eigenvalues. Suppose we have a vanishing linear combination:

$$c_1 \mathbf{v}_1 + \dots + c_{k-1} \mathbf{v}_{k-1} + c_k \mathbf{v}_k = \mathbf{0}. \quad (8.29)$$

Let us multiply this equation by the matrix  $A$ :

$$\begin{aligned} A(c_1 \mathbf{v}_1 + \dots + c_{k-1} \mathbf{v}_{k-1} + c_k \mathbf{v}_k) &= c_1 A \mathbf{v}_1 + \dots + c_{k-1} A \mathbf{v}_{k-1} + c_k A \mathbf{v}_k \\ &= c_1 \lambda_1 \mathbf{v}_1 + \dots + c_{k-1} \lambda_{k-1} \mathbf{v}_{k-1} + c_k \lambda_k \mathbf{v}_k = \mathbf{0}. \end{aligned}$$

On the other hand, if we multiply the original equation (8.29) by  $\lambda_k$ , we also have

$$c_1 \lambda_k \mathbf{v}_1 + \dots + c_{k-1} \lambda_k \mathbf{v}_{k-1} + c_k \lambda_k \mathbf{v}_k = \mathbf{0}.$$

On subtracting this from the previous equation, the final terms cancel, and we are left with the equation

$$c_1(\lambda_1 - \lambda_k) \mathbf{v}_1 + \dots + c_{k-1}(\lambda_{k-1} - \lambda_k) \mathbf{v}_{k-1} = \mathbf{0}.$$

This is a vanishing linear combination of the first  $k - 1$  eigenvectors, and so, by our induction hypothesis, can happen only if all the coefficients are zero:

$$c_1(\lambda_1 - \lambda_k) = 0, \quad \dots \quad c_{k-1}(\lambda_{k-1} - \lambda_k) = 0.$$

The eigenvalues were assumed to be distinct, and consequently  $c_1 = \dots = c_{k-1} = 0$ . Substituting these values back into (8.29), we find that  $c_k \mathbf{v}_k = \mathbf{0}$ , and so  $c_k = 0$  also, since the eigenvector  $\mathbf{v}_k \neq \mathbf{0}$ . Thus we have proved that (8.29) holds if and only if  $c_1 = \dots = c_k = 0$ , which implies the linear independence of the eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_k$ . This completes the induction step. *Q.E.D.*

The most important consequence of this result concerns when a matrix has the maximum allotment of eigenvalues.

**Theorem 8.21.** If the  $n \times n$  real matrix  $A$  has  $n$  distinct real eigenvalues  $\lambda_1, \dots, \lambda_n$ , then the corresponding real eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis of  $\mathbb{R}^n$ . If  $A$  (which may now be either a real or a complex matrix) has  $n$  distinct complex eigenvalues, then the corresponding eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis of  $\mathbb{C}^n$ .

For instance, the  $2 \times 2$  matrix in Example 8.5 has two distinct real eigenvalues, and its two independent eigenvectors form a basis of  $\mathbb{R}^2$ . The  $3 \times 3$  matrix in Example 8.9 has

three distinct complex eigenvalues, and its eigenvectors form a basis for  $\mathbb{C}^3$ . If a matrix has multiple eigenvalues, then there may or may not be an eigenvector basis of  $\mathbb{R}^n$  (or  $\mathbb{C}^n$ ). The matrix in Example 8.6 admits an eigenvector basis, whereas the matrix in Example 8.8 does not. In general, it can be proved<sup>†</sup> that the dimension of an eigenspace is less than or equal to the corresponding eigenvalue's multiplicity. In particular, every simple eigenvalue has a one-dimensional eigenspace, and hence, up to scalar multiple, only one associated eigenvector.

**Definition 8.22.** An eigenvalue  $\lambda$  of a matrix  $A$  is called *complete* if the corresponding eigenspace  $V_\lambda = \ker(A - \lambda I)$  has the same dimension as its multiplicity. The matrix  $A$  is said to be *complete* if all its eigenvalues are complete.

Note that a simple eigenvalue is automatically complete, since its eigenspace is the one-dimensional subspace or eigenline spanned by the corresponding eigenvector. Thus, only multiple eigenvalues can cause a matrix to be incomplete.

**Remark.** The multiplicity of an eigenvalue  $\lambda_i$  is sometimes referred to as its *algebraic multiplicity*. The dimension of the eigenspace  $V_\lambda$  is its *geometric multiplicity*, and so completeness requires that the two multiplicities be equal. The word “complete” is not standard, but has been chosen because it can be used to describe both matrices and their individual eigenvalues. Alternative terms used to describe complete matrices include *perfect*, *semi-simple*, and, as we discuss shortly, *diagonalizable*.

**Theorem 8.23.** An  $n \times n$  real or complex matrix  $A$  is complete if and only if its eigenvectors span  $\mathbb{C}^n$ . In particular, an  $n \times n$  matrix that has  $n$  distinct eigenvalues is complete.

Or, stated another way, a matrix is complete if and only if its eigenvectors form a basis of  $\mathbb{C}^n$ . Most matrices are complete. Incomplete  $n \times n$  matrices, which have fewer than  $n$  linearly independent complex eigenvectors, are less pleasant to deal with, and we relegate most of the messy details to Section 8.6.

**Remark.** We already noted that complex eigenvectors of a real matrix always appear in conjugate pairs:  $\mathbf{v} = \mathbf{x} \pm i\mathbf{y}$ . If the matrix is complete, then it can be shown that its real eigenvectors combined with the real and imaginary parts of its complex conjugate eigenvectors form a real basis for  $\mathbb{R}^n$ . (See Exercise 8.3.12 for the underlying principle.) For instance, the complex eigenvectors of the  $3 \times 3$  matrix appearing in Example 8.9 are

$\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \pm i \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ . The vectors  $\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ ,  $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ , consisting of its real eigenvector

and the real and imaginary parts of its complex eigenvectors, form a basis for  $\mathbb{R}^3$ .

## Exercises

- 8.3.1. Which of the following are complete eigenvalues for the indicated matrix? What is the dimension of the associated eigenspace? (a) 3,  $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ ; (b) 2,  $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ ;

<sup>†</sup> This follows from the Jordan Canonical Form Theorem 8.51.

$$(c) 1, \begin{pmatrix} 0 & 0 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}; \quad (d) 1, \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 2 \end{pmatrix}; \quad (e) -1, \begin{pmatrix} -1 & -4 & -4 \\ 0 & -1 & 0 \\ 0 & 4 & 3 \end{pmatrix};$$

$$(f) -i, \begin{pmatrix} -i & 1 & 0 \\ -i & 1 & -1 \\ 0 & 0 & -i \end{pmatrix}; \quad (g) -2, \begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}; \quad (h) 1, \begin{pmatrix} 1 & -1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 & -1 \\ 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

8.3.2. Find the eigenvalues and a basis for the each of the eigenspaces of the following matrices. Which are complete?

$$(a) \begin{pmatrix} 4 & -4 \\ 1 & 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 6 & -8 \\ 4 & -6 \end{pmatrix}, \quad (c) \begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}, \quad (d) \begin{pmatrix} i & -1 \\ 1 & -i \end{pmatrix}, \quad (e) \begin{pmatrix} 4 & -1 & -1 \\ 0 & 3 & 0 \\ 1 & -1 & 2 \end{pmatrix},$$

$$(f) \begin{pmatrix} -6 & 0 & -8 \\ -4 & 2 & -4 \\ 4 & 0 & 6 \end{pmatrix}, \quad (g) \begin{pmatrix} -2 & 1 & -1 \\ 5 & -3 & 6 \\ 5 & -1 & 4 \end{pmatrix}, \quad (h) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & 0 & -1 & 0 \end{pmatrix}, \quad (i) \begin{pmatrix} -1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ -1 & -4 & 1 & -2 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

8.3.3. Which of the following matrices admit eigenvector bases of  $\mathbb{R}^n$ ? For those that do, exhibit such a basis. If not, what is the dimension of the subspace of  $\mathbb{R}^n$  spanned by the

eigenvectors? (a)  $\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 3 \\ -3 & 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & -2 & 0 \\ 0 & -1 & 0 \\ 4 & -4 & -1 \end{pmatrix}$ ,

$$(e) \begin{pmatrix} 1 & -2 & 0 \\ 0 & -1 & 0 \\ 0 & -4 & -1 \end{pmatrix}, \quad (f) \begin{pmatrix} 2 & 0 & 0 \\ 1 & -1 & 1 \\ 2 & 1 & -1 \end{pmatrix}, \quad (g) \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad (h) \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}.$$

8.3.4. Answer Exercise 8.3.3 with  $\mathbb{R}^n$  replaced by  $\mathbb{C}^n$ .

8.3.5. (a) Give an example of a  $3 \times 3$  matrix with 1 as its only eigenvalue, and only one linearly independent eigenvector. (b) Find one that has two linearly independent eigenvectors.

8.3.6. *True or false:* (a) Every diagonal matrix is complete.  
(b) Every upper triangular matrix is complete.

8.3.7. Prove that if  $A$  is a complete matrix, then so is  $cA + dI$ , where  $c, d$  are any scalars.

8.3.8. (a) Prove that if  $A$  is complete, then so is  $A^2$ .  
(b) Give an example of an incomplete matrix  $A$  such that  $A^2$  is complete.

8.3.9. Let  $U$  be an upper triangular matrix with all its diagonal entries equal. Prove that  $U$  is complete if and only if  $U$  is a diagonal matrix.

◇ 8.3.10. Suppose  $\mathbf{v}_1, \dots, \mathbf{v}_n$  is an eigenvector basis for the complete matrix  $A$ , with  $\lambda_1, \dots, \lambda_n$  the corresponding eigenvalues. Prove that *every* eigenvalue of  $A$  is one of the  $\lambda_1, \dots, \lambda_n$ .

◇ 8.3.11. Show that if  $A$  is complete, then every similar matrix  $B = S^{-1}AS$  is also complete.

◇ 8.3.12. (a) Prove that if  $\mathbf{x} \pm i\mathbf{y}$  is a complex conjugate pair of eigenvectors of a real matrix  $A$  corresponding to complex conjugate eigenvalues  $\mu \pm i\nu$  with  $\nu \neq 0$ , then  $\mathbf{x}$  and  $\mathbf{y}$  are linearly independent real vectors. (b) More generally, if  $\mathbf{v}_j = \mathbf{x}_j \pm i\mathbf{y}_j$ ,  $j = 1, \dots, k$ , are complex conjugate pairs of eigenvectors corresponding to *distinct* pairs of complex conjugate eigenvalues  $\mu_j \pm i\nu_j$ ,  $\nu_j \neq 0$ , then the real vectors  $\mathbf{x}_1, \dots, \mathbf{x}_k, \mathbf{y}_1, \dots, \mathbf{y}_k$  are linearly independent. (c) Prove that if  $A$  is complete, then there exists a basis of  $\mathbb{R}^n$  consisting of its real eigenvectors and real and imaginary parts of its complex eigenvectors.

## Diagonalization

Let  $L: \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a linear transformation on  $n$ -dimensional Euclidean space. As we know, cf. Theorem 7.5,  $L[\mathbf{x}] = A\mathbf{x}$  is prescribed by multiplication by an  $n \times n$  matrix  $A$ . However, the matrix representing a given linear transformation will depend on the choice of basis for the underlying vector space  $\mathbb{R}^n$ . Linear transformations having a complicated matrix representation in terms of the standard basis  $\mathbf{e}_1, \dots, \mathbf{e}_n$  may be considerably simplified by choosing a suitably adapted basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$ . We are now in a position to understand how to effect such a simplification.

For example, the linear transformation  $L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - y \\ 2x + 4y \end{pmatrix}$  studied in Example 7.19 is represented by the matrix  $A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$  — when expressed in terms of the standard basis of  $\mathbb{R}^2$ . In terms of the alternative basis  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ , it is represented by the diagonal matrix  $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ , implying that it has a simple stretching action on the new basis vectors:  $A\mathbf{v}_1 = 2\mathbf{v}_1$ ,  $A\mathbf{v}_2 = 3\mathbf{v}_2$ . Now we can understand the reason for this simplification. *The new basis consists of the two eigenvectors of the matrix  $A$ .* This observation is indicative of a general fact: representing a linear transformation in terms of an eigenvector basis has the effect of changing its matrix representative into a simple diagonal form — thereby *diagonalizing* the original coefficient matrix.

According to (7.31), if  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis of  $\mathbb{R}^n$ , then the corresponding matrix representative of the linear transformation  $L[\mathbf{v}] = A\mathbf{v}$  is given by the similar matrix  $B = S^{-1}AS$ , where  $S = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$  is the matrix whose columns are the basis vectors. In the preceding example,

$$S = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}, \text{ and hence } S^{-1}AS = \begin{pmatrix} 2 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}.$$

**Definition 8.24.** A square matrix  $A$  is called *diagonalizable* if there exists a nonsingular matrix  $S$  and a diagonal matrix  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  such that

$$S^{-1}AS = \Lambda, \quad \text{or, equivalently,} \quad A = S\Lambda S^{-1}. \quad (8.30)$$

A diagonal matrix represents a linear transformation that simultaneously stretches<sup>†</sup> in the direction of the basis vectors. Thus, every diagonalizable matrix represents an elementary combination of (complex) stretching transformations.

To understand the diagonalization equation (8.30), we rewrite it in the equivalent form

$$AS = S\Lambda. \quad (8.31)$$

Using the columnwise action (1.11) of matrix multiplication, one easily sees that the  $k^{\text{th}}$  column of this  $n \times n$  matrix equation is given by

$$A\mathbf{v}_k = \lambda_k\mathbf{v}_k,$$

where  $\mathbf{v}_k$  denotes the  $k^{\text{th}}$  column of  $S$ . Therefore, the columns of  $S$  are necessarily eigenvectors, and the entries of the diagonal matrix  $\Lambda$  are the corresponding eigenvalues. And,

---

<sup>†</sup> A negative diagonal entry represents the combination of a reflection and stretch. Complex entries indicate complex “stretching” transformations. See Section 7.2 for details.

as a result, a diagonalizable matrix  $A$  must have  $n$  linearly independent eigenvectors, i.e., an eigenvector basis, to form the columns of the nonsingular diagonalizing matrix  $S$ . Since the diagonal form  $\Lambda$  contains the eigenvalues along its diagonal, it is uniquely determined up to a permutation of its entries.

Now, as we know, not every matrix has an eigenvector basis. Moreover, even when it exists, the eigenvector basis may be complex, in which case  $S$  is a complex matrix, and the entries of the diagonal matrix  $\Lambda$  are the complex eigenvalues. Thus, we should distinguish between complete matrices that are diagonalizable over the complex numbers and the more restrictive class of real matrices that can be diagonalized by a real matrix  $S$ .

**Theorem 8.25.** A matrix is complex diagonalizable if and only if it is complete. A real matrix is real diagonalizable if and only if it is complete and has all real eigenvalues.

**Example 8.26.** The  $3 \times 3$  matrix  $A = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$  considered in Example 8.5 has eigenvector basis

$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

We assemble these to form the eigenvector matrix

$$S = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \text{whereby} \quad S^{-1} = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The diagonalization equation (8.30) becomes

$$S^{-1} A S = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \Lambda,$$

with eigenvalues of  $A$  appearing on the diagonal of  $\Lambda$ , in the same order as the eigenvectors.

**Remark.** If a matrix is not complete, then it cannot be diagonalized. A simple example is a matrix of the form  $A = \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$  with  $c \neq 0$ , which represents a shear in the direction of the  $x$ -axis. Incomplete matrices represent generalized shearing transformations, and will be the subject of Section 8.6.

## Exercises

- 8.3.13. Diagonalize the following matrices: (a)  $\begin{pmatrix} 3 & -9 \\ 2 & -6 \end{pmatrix}$ , (b)  $\begin{pmatrix} 5 & -4 \\ 2 & -1 \end{pmatrix}$ , (c)  $\begin{pmatrix} -4 & -2 \\ 5 & 2 \end{pmatrix}$ ,  
 (d)  $\begin{pmatrix} -2 & 3 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 3 \end{pmatrix}$ , (e)  $\begin{pmatrix} 8 & 0 & -3 \\ -3 & 0 & -1 \\ 3 & 0 & -2 \end{pmatrix}$ , (f)  $\begin{pmatrix} 3 & 3 & 5 \\ 5 & 6 & 5 \\ -5 & -8 & -7 \end{pmatrix}$ , (g)  $\begin{pmatrix} 2 & 5 & 5 \\ 0 & 2 & 0 \\ 0 & -5 & -3 \end{pmatrix}$ ,  
 (h)  $\begin{pmatrix} 1 & 0 & -1 & 1 \\ 0 & 2 & -1 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$ , (i)  $\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$ , (j)  $\begin{pmatrix} 2 & 1 & -1 & 0 \\ -3 & -2 & 0 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & -1 \end{pmatrix}$ .

8.3.14. Diagonalize the *Fibonacci matrix*  $F = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ .

8.3.15. Diagonalize the matrix  $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$  of rotation through  $90^\circ$ . How would you interpret the result?

8.3.16. Diagonalize the rotation matrices (a)  $\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} \frac{5}{13} & 0 & \frac{12}{13} \\ 0 & 1 & 0 \\ -\frac{12}{13} & 0 & \frac{5}{13} \end{pmatrix}$ .

8.3.17. Which of these matrices have real diagonal forms? (a)  $\begin{pmatrix} -2 & 1 \\ 4 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & 2 \\ -3 & 1 \end{pmatrix}$ ,

(c)  $\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$ , (d)  $\begin{pmatrix} 0 & 3 & 2 \\ -1 & 1 & -1 \\ 1 & -3 & -1 \end{pmatrix}$ , (e)  $\begin{pmatrix} 3 & -8 & 2 \\ -1 & 2 & 2 \\ 1 & -4 & 2 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & -1 \end{pmatrix}$ .

8.3.18. Diagonalize the following complex matrices:

(a)  $\begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}$ , (b)  $\begin{pmatrix} 1-i & 0 \\ i & 2+i \end{pmatrix}$ , (c)  $\begin{pmatrix} 2-i & 2+i \\ 3-i & 1+i \end{pmatrix}$ , (d)  $\begin{pmatrix} -i & 0 & 1 \\ -i & 1 & -1 \\ 1 & 0 & -i \end{pmatrix}$ .

8.3.19. Write down a real matrix that has

(a) eigenvalues  $-1, 3$  and corresponding eigenvectors  $\begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ,

(b) eigenvalues  $0, 2, -2$  and associated eigenvectors  $\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}$ ;

(c) an eigenvalue of  $3$  and corresponding eigenvectors  $\begin{pmatrix} 2 \\ -3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ;

(d) an eigenvalue  $-1+2i$  and corresponding eigenvector  $\begin{pmatrix} 1+i \\ 3i \end{pmatrix}$ ;

(e) an eigenvalue  $-2$  and corresponding eigenvector  $\begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}$ ;

(f) an eigenvalue  $3+i$  and corresponding eigenvector  $\begin{pmatrix} 1 \\ 2i \\ -1-i \end{pmatrix}$ .

8.3.20. A matrix  $A$  has eigenvalues  $-1$  and  $2$  and associated eigenvectors  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$  and  $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$ .

Write down the matrix form of the linear transformation  $L[\mathbf{u}] = A\mathbf{u}$  in terms of (a) the standard basis  $\mathbf{e}_1, \mathbf{e}_2$ ; (b) the basis consisting of its eigenvectors; (c) the basis  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ .

◇ 8.3.21. Prove that two complete matrices  $A, B$  have the same eigenvalues (with multiplicities) if and only if they are similar, i.e.,  $B = S^{-1}AS$  for some nonsingular matrix  $S$ .

8.3.22. Let  $B$  be obtained from  $A$  by permuting both its rows and columns using the same permutation  $\pi$ , so  $b_{ij} = a_{\pi(i),\pi(j)}$ . Prove that  $A$  and  $B$  have the same eigenvalues. How are their eigenvectors related?

8.3.23. *True or false:* If  $A$  is a complete upper triangular matrix, then it has an upper triangular eigenvector matrix  $S$ .

8.3.24. How many different diagonal forms does an  $n \times n$  diagonalizable matrix have?

8.3.25. Characterize all complete matrices that are their own inverses:  $A^{-1} = A$ . Write down a non-diagonal example.

- ◇ 8.3.26. Two  $n \times n$  matrices  $A, B$  are said to be *simultaneously diagonalizable* if there is a nonsingular matrix  $S$  such that both  $S^{-1}AS$  and  $S^{-1}BS$  are diagonal matrices.
- Show that simultaneously diagonalizable matrices commute:  $AB = BA$ .
  - Prove that the converse is valid, provided that one of the matrices has no multiple eigenvalues.
  - Is every pair of commuting matrices simultaneously diagonalizable?

## 8.4 Invariant Subspaces

The notion of an invariant subspace of a linear map plays an important role in dynamical systems, both finite- and infinite-dimensional, as well as in linear iterative systems, and in linear control systems. With the theory of eigenvalues and eigenvectors in hand, we are now able to completely characterize them.

**Definition 8.27.** Let  $L: V \rightarrow V$  be a linear transformation on a vector space  $V$ . A subspace  $W \subset V$  is said to be *invariant* if  $L[\mathbf{w}] \in W$  whenever  $\mathbf{w} \in W$ .

Trivial examples of invariant subspaces, valid for any linear map, are the entire space  $W = V$  and the zero subspace  $W = \{\mathbf{0}\}$ . If  $L = I$  is the identity transformation, then *every* subspace  $W \subset V$  is invariant. More interestingly, both the kernel and image of  $L$  are invariant subspaces. Indeed, when  $\mathbf{w} \in \ker L$ , then  $L[\mathbf{w}] = \mathbf{0} \in \ker L$ , proving invariance. Similarly, if  $\mathbf{w} \in \text{img } L$ , then  $L[\mathbf{w}]$  also lies in  $\text{img } L$  by the definition of image.

**Example 8.28.** Let  $V = \mathbb{R}^2$ . Let us find all invariant subspaces of the scaling transformation  $L(x, y) = (2x, 3y)^T$ . If  $W$  is the line spanned by a vector  $\mathbf{w} = (a, b)^T \neq \mathbf{0}$  then  $L[\mathbf{w}] = (2a, 3b)^T \in W$  if and only if  $(2a, 3b)^T = c\mathbf{w} = (ca, cb)^T$  for some scalar  $c$ . This is clearly possible if and only if either  $a = 0$  or  $b = 0$ . Thus, the only one-dimensional invariant subspaces of this scaling transformation are the  $x$ - and  $y$ -axes.

Next, consider the linear transformation  $L(x, y) = (x + 3y, y)^T$  corresponding to a shear in the direction of the  $x$ -axis. By the same reasoning, the one-dimensional subspace spanned by  $\mathbf{w} = (a, b)^T \neq \mathbf{0}$  is invariant if and only if  $(a + 3b, b)^T = (ca, cb)^T$  for  $c \in \mathbb{R}$ , which is possible only if  $b = 0$ . Thus, the only one-dimensional invariant subspace of this shearing transformation is the  $x$ -axis itself.

Finally, consider the linear transformation  $L(x, y) = (-y, x)^T$  corresponding to counterclockwise rotation by  $90^\circ$ . It is easy to see, either geometrically or algebraically, that  $L$  has no nontrivial invariant subspaces. On the other hand, if we view  $L$  as a map on  $\mathbb{C}^2$ , then one can show that there are two one-dimensional complex invariant subspaces, namely those spanned by its eigenvectors  $\mathbf{w}_1 = (1, i)^T$  and  $\mathbf{w}_2 = (1, -i)^T$ .

From here on, we restrict our attention to the finite-dimensional case, and so the linear transformation  $L$  on either  $\mathbb{R}^n$  or  $\mathbb{C}^n$  is given by matrix multiplication:  $L[\mathbf{x}] = A\mathbf{x}$  for some  $n \times n$  matrix  $A$ .

**Proposition 8.29.** A one-dimensional subspace is invariant under the linear transformation  $L[\mathbf{x}] = A\mathbf{x}$  if and only if it is an eigenline spanned by an eigenvector of  $A$ .

*Proof:* Let  $W$  be spanned by the single non-zero vector  $\mathbf{w} \neq \mathbf{0}$ . Then  $A\mathbf{w} \in W$  if and only if  $A\mathbf{w} = \lambda\mathbf{w}$  for some scalar  $\lambda$ . But this means that  $\mathbf{w}$  is an eigenvector of  $A$  with eigenvalue  $\lambda$ . Q.E.D.

Thus, if  $A$  has no multiple eigenvalues, it has a finite number of one-dimensional invariant subspaces, namely the eigenspaces (eigenlines) associated with each eigenvalue. On the other hand, if  $\lambda$  is a multiple eigenvalue, then every one-dimensional subspace of its eigenspace  $W \subset V_\lambda = \{ \mathbf{v} \mid A\mathbf{v} = \lambda\mathbf{v} \}$  is invariant.

We already observed that if  $A = \lambda I$ , then every subspace  $V \subset \mathbb{C}^n$  is invariant. On the other hand, if  $A$  is diagonal, with all distinct entries, then the invariant subspaces are necessarily spanned by a finite collection of standard basis vectors  $\mathbf{e}_{l_1}, \dots, \mathbf{e}_{l_k}$ , which are recognized as its eigenvectors. This is a special case of the following general characterization of invariant subspaces of complete matrices. The incomplete case will be dealt with in Section 8.6.

**Theorem 8.30.** If  $A$  is a complete matrix, then every  $k$ -dimensional complex invariant subspace is spanned by  $k$  linearly independent eigenvectors of  $A$ .

*Proof:* Let  $W \neq \{\mathbf{0}\}$  be a nontrivial invariant subspace. Thanks to completeness, we can express every nonzero vector  $\mathbf{0} \neq \mathbf{w} \in W$  as a linear combination,  $\mathbf{w} = c_1\mathbf{v}_1 + \dots + c_j\mathbf{v}_j$ , where  $\mathbf{v}_1, \dots, \mathbf{v}_j$  are eigenvectors associated with *distinct* eigenvalues  $\lambda_1, \dots, \lambda_j$  of  $A$  and all coefficients  $c_i \neq 0$ . We claim that this implies that each represented eigenvector  $\mathbf{v}_i \in W$  for  $i = 1, \dots, j$ , which clearly establishes the result. To prove the claim, we write

$$A\mathbf{w} - \lambda_j\mathbf{w} = c_1(\lambda_1 - \lambda_j)\mathbf{v}_1 + \dots + c_{j-1}(\lambda_{j-1} - \lambda_j)\mathbf{v}_{j-1} \in W,$$

since, by the assumption of invariance, both terms on the left hand side belong to  $W$ . Moreover, since the eigenvalues are distinct, we must have  $c_i(\lambda_i - \lambda_j) \neq 0$  for  $i = 1, \dots, j-1$ . Iterating this process, we eventually conclude that a nonzero multiple of  $\mathbf{v}_1$  and hence  $\mathbf{v}_1$  itself belongs to  $W$ . This result is independent of the ordering of the eigenvectors, and hence all  $\mathbf{v}_1, \dots, \mathbf{v}_j \in W$ .  $Q.E.D.$

If  $A$  is a complete real matrix that possesses all real eigenvalues, then the same proof shows that every real invariant subspace has the form given in Theorem 8.30. If  $A$  is real and complete, with complex conjugate eigenvalues, Theorem 8.30 describes its complex invariant subspaces. Its real invariant subspaces are obtained from the real and imaginary parts of the eigenvectors. For example, if  $\mathbf{v}_\pm = \mathbf{x} \pm i\mathbf{y}$  are a complex conjugate pair of eigenvectors, then they individually span one-dimensional complex invariant subspaces. However, the smallest corresponding real invariant subspace is the two-dimensional subspace spanned by  $\mathbf{x}$  and  $\mathbf{y}$ .

**Example 8.31.** Consider the three-dimensional rotation (permutation) matrix

$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ . It has one real eigenvalue,  $\lambda_1 = 1$ , and two complex conjugate eigenvalues,  $\lambda_2 = \frac{1}{2} + \frac{\sqrt{3}}{2}i$  and  $\lambda_3 = \frac{1}{2} - \frac{\sqrt{3}}{2}i$ .

The complex invariant subspaces are spanned by 0, 1, 2, or 3 of the corresponding complex eigenvectors  $(-\frac{1}{2}, -\frac{1}{2}, 1)^T \pm i(\frac{\sqrt{3}}{2}, -\frac{\sqrt{3}}{2}, 0)^T$ . There is a single one-dimensional real invariant subspace, spanned by the real eigenvector  $(1, 1, 1)^T$ , and a single two-dimensional real invariant subspace, which is the orthogonal complement spanned by the real and imaginary parts of the complex conjugate eigenvectors. Indeed, this indicates a general property satisfied by all  $3 \times 3$  rotation matrices, with the exception of the trivial identity matrix. The unique one-dimensional real invariant subspace is the axis of the rotation, and the matrix reduces to a two-dimensional rotation on its orthogonal complement. See Exercise 8.2.44 for further details.

## Exercises

8.4.1. Find all invariant subspaces  $W \subset \mathbb{R}^2$  of the following linear transformations  $L: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ :

- (a) the scaling transformation  $(2x, 3y, 4z)^T$ ;
- (b) the shear  $(x + 3y, y, z)^T$ ;
- (c) counterclockwise rotation by a  $45^\circ$  angle around the  $x$ -axis.

8.4.2. Find all invariant subspaces of the following matrices: (a)  $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 3 & 0 \\ -2 & 3 \end{pmatrix}$ ,

(c)  $\begin{pmatrix} 0 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{pmatrix}$ , (d)  $\begin{pmatrix} -6 & 0 & -8 \\ -4 & 2 & -4 \\ 4 & 0 & 6 \end{pmatrix}$ , (e)  $\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ .

8.4.3. Find all complex invariant subspaces and all real invariant subspaces of

(a)  $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ , (b)  $\begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}$ , (c)  $\begin{pmatrix} 3 & 3 & 5 \\ 5 & 6 & 5 \\ -5 & -8 & -7 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & 5 & 5 \\ 0 & 2 & 0 \\ 0 & -5 & -3 \end{pmatrix}$ .

8.4.4. Prove that if  $W$  is an invariant subspace for  $A$ , then it is also invariant for  $A^2$ . Is the converse to this statement valid?

◇ 8.4.5. Let  $V \subset \mathbb{R}^n$  be an invariant subspace for the  $n \times n$  matrix  $A$ . Explain why every eigenvalue and eigenvector of the linear map obtained by restricting  $A$  to  $V$  are also eigenvalues and eigenvectors of  $A$  itself.

8.4.6. *True or false:* If  $V$  and  $W$  are invariant subspaces for the matrix  $A$ , then so is

- (a)  $V + W$ ;
- (b)  $V \cap W$ ;
- (c)  $V \cup W$ ;
- (d)  $V \setminus W$ .

8.4.7. *True or false:* If  $V$  is an invariant subspace for the  $n \times n$  matrix  $A$  and  $W$  is an invariant subspace for the  $n \times n$  matrix  $B$ , then  $V + W$  is an invariant subspace for the matrix  $A + B$ .

8.4.8. *True or false:* If  $W$  is an invariant subspace of the matrix  $A$ , then it is also an invariant subspace of  $A^T$ .

8.4.9. *True or false:* If  $W$  is an invariant subspace of the nonsingular matrix  $A$ , then it is also an invariant subspace of  $A^{-1}$ .

8.4.10. Which  $2 \times 2$  orthogonal matrices have a nontrivial real invariant subspace?

8.4.11. *True or false:* If  $Q \neq \pm I$  is a  $4 \times 4$  orthogonal matrix, then  $Q$  has no real invariant subspaces.

◇ 8.4.12. (a) Let  $A$  be an  $n \times n$  symmetric matrix, and let  $\mathbf{v}$  be an eigenvector. Prove that its orthogonal complement under the dot product, namely,  $V^\perp = \{\mathbf{w} \in \mathbb{R}^n \mid \mathbf{v}_1 \cdot \mathbf{w} = 0\}$ , is an invariant subspace. (b) More generally, prove that if  $W \subset \mathbb{R}^n$  is an invariant subspace, then its orthogonal complement  $W^\perp$ , is also invariant.

## 8.5 Eigenvalues of Symmetric Matrices

Fortunately, the matrices that arise in most applications are complete and, in fact, possess some additional structure that ameliorates the calculation of their eigenvalues and eigenvectors. The most important class is that of the symmetric, including positive definite, matrices. In fact, not only are the eigenvalues of a symmetric matrix necessarily real, the eigenvectors always form an *orthogonal basis* of the underlying Euclidean space, enjoying all the wonderful properties we studied in Chapter 4. In fact, this is by far the most common way for orthogonal bases to appear — as the eigenvector bases of symmetric matrices. Let us state this important result, but defer its proof until the end of the section.

**Theorem 8.32.** Let  $A = A^T$  be a real symmetric  $n \times n$  matrix. Then

- (a) All the eigenvalues of  $A$  are real.
- (b) Eigenvectors corresponding to distinct eigenvalues are orthogonal.
- (c) There is an orthonormal basis of  $\mathbb{R}^n$  consisting of  $n$  eigenvectors of  $A$ .

In particular, all real symmetric matrices are complete and real diagonalizable.

**Remark.** Orthogonality is with respect to the standard dot product on  $\mathbb{R}^n$ . As we noted in Section 7.5, the transpose is a particular case of the adjoint operation when we use the Euclidean dot product. An analogous result holds for more general self-adjoint linear transformations under more general inner products on  $\mathbb{R}^n$ ; see Exercise 8.5.10.

**Example 8.33.** The  $2 \times 2$  matrix  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$  considered in Example 8.5 is symmetric, and so has real eigenvalues  $\lambda_1 = 4$  and  $\lambda_2 = 2$ . You can easily check that the corresponding eigenvectors  $\mathbf{v}_1 = (1, 1)^T$  and  $\mathbf{v}_2 = (-1, 1)^T$  are orthogonal:  $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$ , and hence form an orthogonal basis of  $\mathbb{R}^2$ . The orthonormal eigenvector basis promised by Theorem 8.32 is obtained by dividing each eigenvector by its Euclidean norm:

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}.$$

**Example 8.34.** Consider the symmetric matrix  $A = \begin{pmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & -1 \end{pmatrix}$ . A straightforward computation produces its eigenvalues and eigenvectors:

$$\begin{aligned} \lambda_1 &= 9, & \lambda_2 &= 3, & \lambda_3 &= -3, \\ \mathbf{v}_1 &= \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}. \end{aligned}$$

As the reader can check, the eigenvectors form an orthogonal basis of  $\mathbb{R}^3$ . An orthonormal basis is provided by the corresponding unit eigenvectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{pmatrix}.$$

The eigenvalues of a symmetric matrix can be used to test its positive definiteness.

**Theorem 8.35.** A symmetric matrix  $K = K^T$  is positive definite if and only if all of its eigenvalues are strictly positive.

*Proof:* First, if  $K > 0$ , then, by definition,  $\mathbf{x}^T K \mathbf{x} > 0$  for all nonzero vectors  $\mathbf{x} \in \mathbb{R}^n$ . In particular, if  $\mathbf{x} = \mathbf{v} \neq \mathbf{0}$  is an eigenvector with (necessarily real) eigenvalue  $\lambda$ , then

$$0 < \mathbf{v}^T K \mathbf{v} = \mathbf{v}^T (\lambda \mathbf{v}) = \lambda \mathbf{v}^T \mathbf{v} = \lambda \|\mathbf{v}\|^2, \quad (8.32)$$

which immediately proves that  $\lambda > 0$ .

Conversely, suppose  $K$  has all positive eigenvalues. Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be the orthonormal eigenvector basis guaranteed by Theorem 8.32, with  $K\mathbf{u}_j = \lambda_j \mathbf{u}_j$  with  $\lambda_j > 0$ . Writing

$$\mathbf{x} = c_1 \mathbf{u}_1 + \dots + c_n \mathbf{u}_n, \quad \text{we obtain} \quad K\mathbf{x} = c_1 \lambda_1 \mathbf{u}_1 + \dots + c_n \lambda_n \mathbf{u}_n.$$

Therefore, using the orthonormality of the eigenvectors,

$$\mathbf{x}^T K \mathbf{x} = (c_1 \mathbf{u}_1^T + \dots + c_n \mathbf{u}_n^T) (c_1 \lambda_1 \mathbf{u}_1 + \dots + c_n \lambda_n \mathbf{u}_n) = \lambda_1 c_1^2 + \dots + \lambda_n c_n^2 > 0$$

whenever  $\mathbf{x} \neq \mathbf{0}$ , since only  $\mathbf{x} = \mathbf{0}$  has coordinates  $c_1 = \dots = c_n = 0$ . This establishes the positive definiteness of  $K$ . *Q.E.D.*

**Remark.** The same proof shows that  $K$  is positive semi-definite if and only if all its eigenvalues satisfy  $\lambda \geq 0$ . A positive semi-definite matrix that is not positive definite admits a zero eigenvalue and one or more *null eigenvectors*, i.e., solutions to  $K\mathbf{v} = \mathbf{0}$ . Every nonzero element  $\mathbf{0} \neq \mathbf{v} \in \ker K$  of its kernel is a null eigenvector.

**Example 8.36.** The symmetric matrix  $K = \begin{pmatrix} 8 & 0 & 1 \\ 0 & 8 & 1 \\ 1 & 1 & 7 \end{pmatrix}$  has characteristic equation

$$\det(K - \lambda I) = -\lambda^3 + 23\lambda^2 - 174\lambda + 432 = -(\lambda - 9)(\lambda - 8)(\lambda - 6),$$

and so its eigenvalues are 9, 8, and 6. Since they are all positive,  $K$  is a positive definite matrix. The associated eigenvectors are

$$\lambda_1 = 9, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = 8, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad \lambda_3 = 6, \quad \mathbf{v}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

Note that the eigenvectors form an orthogonal basis of  $\mathbb{R}^3$ , as guaranteed by Theorem 8.32. As usual, we can construct an orthonormal eigenvector basis

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{pmatrix},$$

by dividing each eigenvector by its norm.

*Proof of Theorem 8.32:* First recall that (see Exercise 3.6.38) if  $A = A^T$  is real, symmetric, then

$$(A\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\mathbf{w}) \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{C}^n, \quad (8.33)$$

where  $\cdot$  indicates the Euclidean dot product when the vectors are real and, more generally, the Hermitian dot product  $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \overline{\mathbf{w}}$  when they are complex.

To prove property (a), suppose  $\lambda$  is a complex eigenvalue with complex eigenvector  $\mathbf{v} \in \mathbb{C}^n$ . Consider the Hermitian dot product of the complex vectors  $A\mathbf{v}$  and  $\mathbf{v}$ :

$$(A\mathbf{v}) \cdot \mathbf{v} = (\lambda\mathbf{v}) \cdot \mathbf{v} = \lambda \|\mathbf{v}\|^2.$$

On the other hand, by (8.33),

$$(A\mathbf{v}) \cdot \mathbf{v} = \mathbf{v} \cdot (A\mathbf{v}) = \mathbf{v} \cdot (\lambda\mathbf{v}) = \mathbf{v}^T \overline{\lambda\mathbf{v}} = \overline{\lambda} \|\mathbf{v}\|^2.$$

Equating these two expressions, we deduce

$$\bar{\lambda} \|\mathbf{v}\|^2 = \lambda \|\mathbf{v}\|^2.$$

Since  $\mathbf{v}$  is an eigenvector, it must be nonzero. Thus, we deduce that  $\bar{\lambda} = \lambda$ , proving that the eigenvalue  $\lambda$  must be real.

To prove (b), suppose  $A\mathbf{v} = \lambda\mathbf{v}$ ,  $A\mathbf{w} = \mu\mathbf{w}$ , where  $\lambda \neq \mu$  are distinct real eigenvalues. Then, again by (8.33),

$$\lambda \mathbf{v} \cdot \mathbf{w} = (A\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\mathbf{w}) = \mathbf{v} \cdot (\mu\mathbf{w}) = \mu \mathbf{v} \cdot \mathbf{w}, \quad \text{and hence } (\lambda - \mu) \mathbf{v} \cdot \mathbf{w} = 0.$$

Since  $\lambda \neq \mu$ , this implies that  $\mathbf{v} \cdot \mathbf{w} = 0$ , so the eigenvectors  $\mathbf{v}, \mathbf{w}$  are orthogonal.

Finally, the proof of (c) is easy if all the eigenvalues of  $A$  are distinct. Theorem 8.21 implies that the eigenvectors form a basis of  $\mathbb{R}^n$ , and part (b) proves they are orthogonal. (An alternative proof starts with orthogonality, and then applies Proposition 4.4 to prove that the eigenvectors form a basis.) To obtain an orthonormal basis, we merely divide the eigenvectors by their lengths:  $\mathbf{u}_k = \mathbf{v}_k / \|\mathbf{v}_k\|$ , as in Lemma 4.2.

To prove (c) in general, we proceed by induction on the size  $n$  of the matrix  $A$ . To start, the case of a  $1 \times 1$  matrix is trivial. (Why?) Next, suppose  $A$  has size  $n \times n$ . We know that  $A$  has at least one eigenvalue,  $\lambda_1$ , which is necessarily real. Let  $\mathbf{v}_1$  be an associated eigenvector. Let  $V^\perp = \{\mathbf{w} \in \mathbb{R}^n \mid \mathbf{v}_1 \cdot \mathbf{w} = 0\}$  denote its orthogonal complement — the subspace of all vectors orthogonal to the first eigenvector. Proposition 4.41 implies that  $\dim V^\perp = n-1$ , and so we can choose an orthonormal basis  $\mathbf{y}_1, \dots, \mathbf{y}_{n-1}$ . Moreover, by Exercise 8.4.12,  $V^\perp$  is an invariant subspace of  $A$ , and hence  $A$  defines a linear transformation on  $V^\perp$  that is represented by an  $(n-1) \times (n-1)$  matrix, say  $B = (b_{ij})$ ,  $i, j = 1, \dots, n-1$ , with respect to the chosen orthonormal basis  $\mathbf{y}_1, \dots, \mathbf{y}_{n-1}$ , so that  $A\mathbf{y}_i = \sum_{j=1}^{n-1} b_{ij} \mathbf{y}_j$ . Moreover, by orthonormality and (8.33),

$$b_{ij} = \mathbf{y}_i \cdot A\mathbf{y}_j = (A\mathbf{y}_i) \cdot \mathbf{y}_j = b_{ji},$$

and hence  $B = B^T$  is symmetric. Our induction hypothesis then implies that there is an orthonormal basis of  $V^\perp \subset \mathbb{R}^n$  consisting of eigenvectors  $\mathbf{u}_2, \dots, \mathbf{u}_n$  of  $B$ , and hence also of  $A$ , each of which is orthogonal to  $\mathbf{v}_1$ . Appending the unit eigenvector  $\mathbf{u}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|$  to this collection will complete the orthonormal basis of  $\mathbb{R}^n$ . Q.E.D.

**Proposition 8.37.** Let  $A = A^T$  be an  $n \times n$  symmetric matrix. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  be an orthogonal eigenvector basis such that  $\mathbf{v}_1, \dots, \mathbf{v}_r$  correspond to nonzero eigenvalues, while  $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$  are null eigenvectors corresponding to the zero eigenvalue (if any). Then  $r = \text{rank } A$ ; the non-null eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_r$  form an orthogonal basis for  $\text{img } A = \text{coimg } A$ , while the null eigenvectors  $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$  form an orthogonal basis for  $\ker A = \text{coker } A$ .

*Proof:* The zero eigenspace coincides with the kernel,  $V_0 = \ker A$ . Thus, the linearly independent null eigenvectors form a basis for  $\ker A$ , which has dimension  $n - r$  where  $r = \text{rank } A$ . Moreover, the remaining  $r$  non-null eigenvectors are orthogonal to the null eigenvectors. Therefore, they must form a basis for the kernel's orthogonal complement, namely  $\text{coimg } A = \text{img } A$ . Q.E.D.

## Exercises

8.5.1. Find the eigenvalues and an orthonormal eigenvector basis for the following symmetric matrices:

$$(a) \begin{pmatrix} 2 & 6 \\ 6 & -7 \end{pmatrix}, (b) \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix}, (c) \begin{pmatrix} 2 & -1 \\ -1 & 5 \end{pmatrix}, (d) \begin{pmatrix} 1 & 0 & 4 \\ 0 & 1 & 3 \\ 4 & 3 & 1 \end{pmatrix}, (e) \begin{pmatrix} 6 & -4 & 1 \\ -4 & 6 & -1 \\ 1 & -1 & 11 \end{pmatrix}.$$

8.5.2. Determine whether the following symmetric matrices are positive definite by computing their eigenvalues. Validate your conclusions by using the methods from Chapter 5.

$$(a) \begin{pmatrix} 2 & -2 \\ -2 & 3 \end{pmatrix}, (b) \begin{pmatrix} -2 & 3 \\ 3 & 6 \end{pmatrix}, (c) \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, (d) \begin{pmatrix} 4 & -1 & -2 \\ -1 & 4 & -1 \\ -2 & -1 & 4 \end{pmatrix}.$$

8.5.3. Prove that a symmetric matrix is negative definite if and only if all its eigenvalues are negative.

8.5.4. How many orthonormal eigenvector bases does a symmetric  $n \times n$  matrix have?

8.5.5. Let  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ . (a) Write down necessary and sufficient conditions on the entries  $a, b, c, d$  that ensures that  $A$  has only real eigenvalues.

(b) Verify that all symmetric  $2 \times 2$  matrices satisfy your conditions.

(c) Write down a non-symmetric matrix that satisfies your conditions.

◇ 8.5.6. Let  $A^T = -A$  be a real, skew-symmetric  $n \times n$  matrix. (a) Prove that the only possible real eigenvalue of  $A$  is  $\lambda = 0$ . (b) More generally, prove that all eigenvalues  $\lambda$  of  $A$  are purely imaginary, i.e.,  $\operatorname{Re} \lambda = 0$ . (c) Explain why 0 is an eigenvalue of  $A$  whenever  $n$  is odd. (d) Explain why, if  $n = 3$ , the eigenvalues of  $A \neq 0$  are  $0, i\omega, -i\omega$ , for some real  $\omega \neq 0$ . (e) Verify these facts for the particular matrices

$$(i) \begin{pmatrix} 0 & -2 \\ 2 & 0 \end{pmatrix}, (ii) \begin{pmatrix} 0 & 3 & 0 \\ -3 & 0 & -4 \\ 0 & 4 & 0 \end{pmatrix}, (iii) \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}, (iv) \begin{pmatrix} 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -3 \\ -2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}.$$

◇ 8.5.7. (a) Prove that every eigenvalue of a Hermitian matrix  $A$ , satisfying  $A^T = \bar{A}$  as in Exercise 3.6.45, is real. (b) Show that the eigenvectors corresponding to distinct eigenvalues are orthogonal under the Hermitian dot product on  $\mathbb{C}^n$ . (c) Find the eigenvalues and eigenvectors of the following Hermitian matrices, and verify orthogonality:

$$(i) \begin{pmatrix} 2 & i \\ -i & -2 \end{pmatrix}, (ii) \begin{pmatrix} 3 & 2-i \\ 2+i & -1 \end{pmatrix}, (iii) \begin{pmatrix} 0 & i & 0 \\ -i & 0 & i \\ 0 & -i & 0 \end{pmatrix}.$$

◇ 8.5.8. Let  $K, M$  be  $n \times n$  matrices, with  $M > 0$  positive definite. A nonzero vector  $\mathbf{v} \neq \mathbf{0}$  is called a *generalized eigenvector* of matrix pair  $K, M$  if it satisfies the *generalized eigenvalue equation*

$$K\mathbf{v} = \lambda M\mathbf{v}, \quad \mathbf{v} \neq \mathbf{0}, \quad (8.34)$$

where the scalar  $\lambda$  is the corresponding *generalized eigenvalue*. Note that ordinary eigenvalue/eigenvectors of  $K$  are when  $M = I$ . (a) Prove that  $\lambda$  is a generalized eigenvalue if and only if it satisfies the *generalized characteristic equation*  $\det(K - \lambda M) = 0$ .

(b) Prove that  $\lambda$  is a generalized eigenvalue of the matrix pair  $K, M$  if and only if it is an ordinary eigenvalue of the matrix  $M^{-1}K$ . How are the eigenvectors related? (c) Now suppose  $K$  is a symmetric matrix. Prove that its generalized eigenvalues are all real.

*Hint:* First explain why this does *not* follow from part (a). Instead mimic the proof of part (a) of Theorem 8.32, using the weighted Hermitian inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T M \bar{\mathbf{w}}$  in place of the dot product. (d) Show that if  $K > 0$ , then its generalized eigenvalues are all positive:  $\lambda > 0$ . (e) Prove that the eigenvectors corresponding to different generalized eigenvalues are orthogonal under the weighted inner product  $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^T M \bar{\mathbf{w}}$ . (f) Show

that, if the matrix pair  $K, M$  has  $n$  distinct generalized eigenvalues, then the eigenvectors form an orthogonal basis for  $\mathbb{R}^n$ . **Remark.** One can, by mimicking the proof of part (c) of Theorem 8.32, show that this holds even when there are repeated generalized eigenvalues.

8.5.9. Compute the generalized eigenvalues and eigenvectors, as in (8.34), for the following matrix pairs. Verify orthogonality of the eigenvectors under the appropriate inner product.

- (a)  $K = \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$ ,  $M = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ , (b)  $K = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$ ,  $M = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ ,  
 (c)  $K = \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}$ ,  $M = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$ , (d)  $K = \begin{pmatrix} 6 & -8 & 3 \\ -8 & 24 & -6 \\ 3 & -6 & 99 \end{pmatrix}$ ,  $M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix}$ ,  
 (e)  $K = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 8 & 2 \\ 0 & 2 & 1 \end{pmatrix}$ ,  $M = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ , (f)  $K = \begin{pmatrix} 5 & 3 & -5 \\ 3 & 3 & -1 \\ -5 & -1 & 9 \end{pmatrix}$ ,  $M = \begin{pmatrix} 3 & 2 & -3 \\ 2 & 2 & -1 \\ -3 & -1 & 5 \end{pmatrix}$ .

◇ 8.5.10. Let  $L = L^* : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a self-adjoint linear transformation with respect to the inner product  $\langle \cdot, \cdot \rangle$ . Prove that all its eigenvalues are real and the eigenvectors are orthogonal. *Hint:* Mimic the proof of Theorem 8.32, replacing the dot product by the given inner product.

◇ 8.5.11. The *difference map*  $\Delta : \mathbb{C}^n \rightarrow \mathbb{C}^n$  is defined as  $\Delta = S - I$ , where  $S$  is the shift map of Exercise 8.2.13. (a) Write down the matrix  $D$  corresponding to  $\Delta$ . (b) Prove that the sampled exponential vectors  $\omega_0, \dots, \omega_{n-1}$  from (5.102) form an eigenvector basis of  $D$ . What are the eigenvalues? (c) Prove that  $K = D^T D$  has the same eigenvectors as  $D$ . What are its eigenvalues? (d) Is  $K$  positive definite? (e) According to Theorem 8.32 the eigenvectors of a symmetric matrix are real and orthogonal. Use this to explain the orthogonality of the sampled exponential vectors. But, why aren't they real?

◇ 8.5.12. An  $n \times n$  *circulant matrix* has the form  $C = \begin{pmatrix} c_0 & c_1 & c_2 & c_3 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & c_2 & \cdots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & c_1 & \cdots & c_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & c_4 & \cdots & c_0 \end{pmatrix}$ ,

in which the entries of each succeeding row are obtained by moving all the previous row's entries one slot to the right, the last entry moving to the front. (a) Check that the shift matrix  $A$  of Exercise 8.2.13, the difference matrix  $D$ , and its symmetric product  $K = D^T D$  of Exercise 8.5.11 are all circulant matrices. (b) Prove that the sampled exponential vectors  $\omega_0, \dots, \omega_{n-1}$ , cf. (5.102), are eigenvectors of  $C$ . Thus, all circulant matrices have *the same eigenvectors!* What are the eigenvalues? (c) Prove that  $F_n^{-1} C F_n = \Lambda$ , where  $F_n$  is the Fourier matrix in Exercise 5.6.9 and  $\Lambda$  is the diagonal matrix with the eigenvalues of  $C$  along the diagonal. (d) Find the eigenvalues and eigenvectors of the following circulant matrices:

$$(i) \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad (ii) \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}, \quad (iii) \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 \end{pmatrix}, \quad (iv) \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}.$$

(e) Find the eigenvalues of the tricirculant matrices in Exercise 1.7.13. Can you find a general formula for the  $n \times n$  version? Explain why the eigenvalues must be real and positive. Does your formula reflect this fact? (f) Which of the preceding matrices are invertible? Write down a general criterion for checking the invertibility of circulant matrices.

## The Spectral Theorem

Every real, symmetric matrix admits an eigenvector basis, and hence is diagonalizable. Moreover, since we can choose eigenvectors that form an orthonormal basis, the diagonalizing matrix takes a particularly simple form. Recall that an  $n \times n$  matrix  $Q$  is *orthogonal* if and only if its columns form an orthonormal basis of  $\mathbb{R}^n$ . Alternatively, one characterizes orthogonal matrices by the condition  $Q^{-1} = Q^T$ , as in Definition 4.18.

Using the orthonormal eigenvector basis in the diagonalization formula (8.30) results in what is known as the *spectral factorization* of a symmetric matrix.

**Theorem 8.38.** Let  $A$  be a real, symmetric matrix. Then there exists an orthogonal matrix  $Q$  such that

$$A = Q \Lambda Q^{-1} = Q \Lambda Q^T, \quad (8.35)$$

where  $\Lambda$  is a real diagonal matrix. The eigenvalues of  $A$  appear on the diagonal of  $\Lambda$ , while the columns of  $Q$  are the corresponding orthonormal eigenvectors.

**Remark.** The term “spectrum” refers to the eigenvalues of a matrix, or, more generally, a linear operator. The terminology is motivated by physics. The spectral energy lines of atoms, molecules, and nuclei are characterized as the eigenvalues of the governing quantum mechanical Schrödinger operator, [54]. The *Spectral Theorem* 8.38 is the finite-dimensional version of the decomposition of quantum mechanical linear operators into their spectral eigenstates.

**Warning.** Although both involve diagonal matrices, the spectral factorization  $A = Q \Lambda Q^T$  and the Gaussian factorization  $A = LDL^T$  of a regular symmetric matrix, cf. (1.58), are completely different. In particular, the eigenvalues are *not* the pivots, so  $\Lambda \neq D$ .

The spectral factorization (8.35) provides us with an alternative means of diagonalizing the associated quadratic form  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$ , i.e., of completing the square. We write

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} = \mathbf{x}^T Q \Lambda Q^T \mathbf{x} = \mathbf{y}^T \Lambda \mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2, \quad (8.36)$$

where the entries of  $\mathbf{y} = Q^T \mathbf{x} = Q^{-1} \mathbf{x}$  are the coordinates of  $\mathbf{x}$  with respect to the orthonormal eigenvector basis of  $A$ . In particular,  $q(\mathbf{x}) > 0$  for all  $\mathbf{x} \neq \mathbf{0}$  and so  $A$  is positive definite if and only if each eigenvalue  $\lambda_i$  is strictly positive, reconfirming Theorem 8.35.

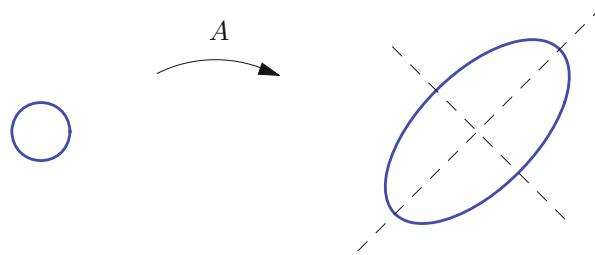
**Example 8.39.** For the  $2 \times 2$  matrix  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$  considered in Example 8.33, the orthonormal eigenvectors produce the diagonalizing orthogonal matrix  $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ . The reader can validate the resulting spectral factorization:

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = A = Q \Lambda Q^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

According to formula (8.36), the associated quadratic form is diagonalized as

$$q(\mathbf{x}) = 3x_1^2 + 2x_1 x_2 + 3x_2^2 = 4y_1^2 + 2y_2^2,$$

where  $\mathbf{y} = Q^T \mathbf{x}$ , i.e.,  $y_1 = \frac{x_1 + x_2}{\sqrt{2}}$ ,  $y_2 = \frac{-x_1 + x_2}{\sqrt{2}}$ .



**Figure 8.3.** Stretching a Circle into an Ellipse.

You can always choose  $Q$  to be a proper orthogonal matrix, so  $\det Q = 1$ , since an improper orthogonal matrix can be made proper by multiplying one of its columns by  $-1$ , which does not affect its status as an eigenvector matrix. Since a proper orthogonal matrix  $Q$  represents a rigid rotation of  $\mathbb{R}^n$ , the diagonalization of a symmetric matrix can be interpreted as a rotation of the coordinate system that makes the orthogonal eigenvectors line up along the coordinate axes. Therefore, a linear transformation  $L[\mathbf{x}] = A\mathbf{x}$  represented by a positive definite matrix  $A$  can be regarded as a combination of stretches in  $n$  mutually orthogonal directions. A good way to visualize this is to consider the effect of the linear transformation on the unit (Euclidean) sphere  $S_1 = \{ \|\mathbf{x}\| = 1 \}$ . Stretching the sphere in mutually orthogonal directions will map it to an ellipsoid  $E = L[S_1] = \{ A\mathbf{x} \mid \|\mathbf{x}\| = 1 \}$  whose principal axes are aligned with the directions of stretch; see Figure 8.3 for the two-dimensional case. For instance, in elasticity, the stress tensor of a deformed body is represented by a positive definite matrix. Its eigenvalues are known as the *principal stretches* and its eigenvectors the *principal directions* of the elastic deformation.

## Exercises

8.5.13. Write out the spectral factorization of the following matrices:

$$(a) \begin{pmatrix} -3 & 4 \\ 4 & 3 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}.$$

8.5.14. Write out the spectral factorization of the matrices listed in Exercise 8.5.1.

8.5.15. Construct a symmetric matrix with the following eigenvectors and eigenvalues, or

explain why none exists: (a)  $\lambda_1 = 1$ ,  $\mathbf{v}_1 = \left( \frac{3}{5}, \frac{4}{5} \right)^T$ ,  $\lambda_2 = 3$ ,  $\mathbf{v}_2 = \left( -\frac{4}{5}, \frac{3}{5} \right)^T$ ,  
 (b)  $\lambda_1 = -2$ ,  $\mathbf{v}_1 = (1, -1)^T$ ,  $\lambda_2 = 1$ ,  $\mathbf{v}_2 = (1, 1)^T$ , (c)  $\lambda_1 = 3$ ,  $\mathbf{v}_1 = (2, -1)^T$ ,  
 $\lambda_2 = -1$ ,  $\mathbf{v}_2 = (-1, 2)^T$ , (d)  $\lambda_1 = 2$ ,  $\mathbf{v}_1 = (2, 1)^T$ ,  $\lambda_2 = 2$ ,  $\mathbf{v}_2 = (1, 2)^T$ .

8.5.16. (a) Find the eigenvalues and eigenvectors of the matrix  $A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}$ .

(b) Use the eigenvalues to compute the determinant of  $A$ . (c) Is  $A$  positive definite? Why or why not? (d) Find an orthonormal eigenvector basis of  $\mathbb{R}^3$  determined by  $A$  or explain why none exists. (e) Write out the spectral factorization of  $A$  if possible. (f) Use orthogonality to write the vector  $(1, 0, 0)^T$  as a linear combination of eigenvectors of  $A$ .

8.5.17. Use the spectral factorization to diagonalize the following quadratic forms:

$$(a) x^2 - 3xy + 5y^2, \quad (b) 3x^2 + 4xy + 6y^2, \quad (c) x^2 + 8xz + y^2 + 6yz + z^2, \\ (d) \frac{3}{2}x^2 - xy - xz + y^2 + z^2, \quad (e) 6x^2 - 8xy + 2xz + 6y^2 - 2yz + 11z^2.$$

◇ 8.5.18. Let  $\mathbf{u}_1, \dots, \mathbf{u}_n$  be an orthonormal basis of  $\mathbb{R}^n$ . Prove that it forms an eigenvector basis for some symmetric  $n \times n$  matrix  $A$ . Can you characterize all such matrices?

8.5.19. *True or false:* A matrix with a real orthonormal eigenvector basis is symmetric.

8.5.20. Prove that every quadratic form can be written as  $\mathbf{x}^T A \mathbf{x} = \|\mathbf{x}\|^2 \left( \sum_{i=1}^n \lambda_i \cos^2 \theta_i \right)$ ,

where  $\lambda_i$  are the eigenvalues of  $A$  and  $\theta_i = \hat{\chi}(\mathbf{x}, \mathbf{v}_i)$  denotes the angle between  $\mathbf{x}$  and the  $i^{\text{th}}$  eigenvector.

8.5.21. An elastic body has stress tensor  $T = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix}$ . Find the principal stretches and principal directions of stretch.

◇ 8.5.22. Given a solid body spinning around its center of mass, the eigenvectors of its positive definite *inertia tensor* prescribe three mutually orthogonal *principal directions* of rotation, while the corresponding eigenvalues are the *moments of inertia*. Given the inertia tensor  $T = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ , find the principal directions and moments of inertia.

◇ 8.5.23. Let  $K$  be a positive definite  $2 \times 2$  matrix. (a) Explain why the quadratic equation  $\mathbf{x}^T K \mathbf{x} = 1$  defines an ellipse. Prove that its principal axes are the eigenvectors of  $K$ , and the semi-axes are the reciprocals of the square roots of the eigenvalues.

(b) Graph and describe the following curves:

$$(i) x^2 + 4y^2 = 1, \quad (ii) x^2 + xy + y^2 = 1, \quad (iii) 3x^2 + 2xy + y^2 = 1.$$

(c) What sort of curve(s) does  $\mathbf{x}^T K \mathbf{x} = 1$  describe if  $K$  is not positive definite?

◇ 8.5.24. Let  $K$  be a positive definite  $3 \times 3$  matrix. (a) Prove that the quadratic equation  $\mathbf{x}^T K \mathbf{x} = 1$  defines an ellipsoid in  $\mathbb{R}^3$ . What are its principal axes and semi-axes?

(b) Describe the surface defined by the equation  $11x^2 - 8xy + 20y^2 - 10xz + 8yz + 11z^2 = 1$ .

8.5.25. Prove that  $A = A^T$  has a repeated eigenvalue if and only if it commutes,  $AJ = JA$ , with a nonzero skew-symmetric matrix:  $J^T = -J \neq 0$ .

*Hint:* First prove this when  $A$  is a diagonal matrix.

8.5.26. Find all positive definite orthogonal matrices.

◇ 8.5.27. (a) Prove that every positive definite matrix  $K$  has a unique positive definite *square root*, i.e., a matrix  $B > 0$  satisfying  $B^2 = K$ .

(b) Find the positive definite square roots of the following matrices:

$$(i) \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad (ii) \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}, \quad (iii) \begin{pmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{pmatrix}, \quad (iv) \begin{pmatrix} 6 & -4 & 1 \\ -4 & 6 & -1 \\ 1 & -1 & 11 \end{pmatrix}.$$

◇ 8.5.28. *The Polar Decomposition:* Prove that every invertible matrix  $A$  has a *polar decomposition*, written  $A = QB$ , into the product of an orthogonal matrix  $Q$  and a positive definite matrix  $B > 0$ . Show that if  $\det A > 0$ , then  $Q$  is a proper orthogonal matrix. *Hint:* Look at the Gram matrix  $K = A^T A$  and use Exercise 8.5.27.

**Remark.** In mechanics, if  $A$  represents the deformation of a body, then  $Q$  represents a rotation, while  $B$  represents a stretching along the orthogonal eigendirections of  $K$ . Thus, every linear deformation of an elastic body can be decomposed into a pure stretching transformation followed by a rotation.

8.5.29. Find the polar decompositions  $A = QB$ , as defined in Exercise 8.5.28, of the following matrices:

$$(a) \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & -3 \\ 1 & 6 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 0 & -3 & 8 \\ 1 & 0 & 0 \\ 0 & 4 & 6 \end{pmatrix}, \quad (e) \begin{pmatrix} 1 & 0 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

◇ 8.5.30. *The Spectral Theorem for Hermitian Matrices.* Prove that a complex Hermitian matrix can be factored as  $H = U \Lambda U^\dagger$  where  $U$  is a unitary matrix and  $\Lambda$  is a real diagonal matrix. *Hint:* See Exercises 4.3.25, 8.5.7.

8.5.31. Find the spectral factorization, as in Exercise 8.5.30, of the following Hermitian matrices:

$$(a) \begin{pmatrix} 3 & 2i \\ -2i & 6 \end{pmatrix}, \quad (b) \begin{pmatrix} 6 & 1-2i \\ 1+2i & 2 \end{pmatrix}, \quad (c) \begin{pmatrix} -1 & 5i & -4 \\ -5i & -1 & 4i \\ -4 & -4i & 8 \end{pmatrix}.$$

◇ 8.5.32. *The Spectral Decomposition:* Let  $A$  be a symmetric matrix with distinct eigenvalues  $\lambda_1, \dots, \lambda_k$ . Let  $V_j = \ker(A - \lambda_j I)$  denote the eigenspace corresponding to  $\lambda_j$ , and let  $P_j$  be the orthogonal projection matrix onto  $V_j$ , as defined in Exercise 4.4.9. (i) Prove that the spectral factorization (8.35) can be rewritten as

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \dots + \lambda_k P_k, \quad (8.37)$$

expressing  $A$  as a linear combination of projection matrices. (ii) Write out the *spectral decomposition* (8.37) for the matrices in Exercise 8.5.13. (iii) Show that

$$I = P_1 + P_2 + \dots + P_k, \quad \text{while} \quad P_i^2 = P_i, \quad P_i P_j = O \quad \text{for } i \neq j.$$

(iv) Show that if  $p(t)$  is any polynomial, then

$$p(A) = p(\lambda_1) P_1 + p(\lambda_2) P_2 + \dots + p(\lambda_k) P_k. \quad (8.38)$$

**Remark.** Replacing  $p(t)$  by any function  $f(t)$  allows one to define  $f(A)$  for any symmetric matrix  $A$ .

## Optimization Principles for Eigenvalues of Symmetric Matrices

As we learned in Chapter 5, the solution to a linear system with positive definite coefficient matrix can be characterized by a minimization principle. Thus, it should come as no surprise that eigenvalues of positive definite matrices, and even more general symmetric matrices, can also be characterized by some sort of optimization procedure. A number of basic numerical algorithms for computing eigenvalues of matrices are based on such optimization principles.

First, consider the relatively simple case of a real diagonal matrix  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ . We assume that the diagonal entries, which are the *same* as the eigenvalues, appear in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n, \quad (8.39)$$

so  $\lambda_1$  is the largest eigenvalue, while  $\lambda_n$  is the smallest. The effect of  $\Lambda$  on a vector  $\mathbf{y} = (y_1, y_2, \dots, y_n)^T \in \mathbb{R}^n$  is to multiply its entries by the diagonal eigenvalues:  $\Lambda \mathbf{y} = (\lambda_1 y_1, \lambda_2 y_2, \dots, \lambda_n y_n)^T$ . In other words, the linear transformation represented by the coefficient matrix  $\Lambda$  has the effect of stretching<sup>†</sup> the  $i^{\text{th}}$  coordinate direction by the factor  $\lambda_i$ . In particular, the maximal stretch occurs in the  $\mathbf{e}_1$  direction, with factor  $\lambda_1$ , while the minimal (or largest negative) stretch occurs in the  $\mathbf{e}_n$  direction, with factor  $\lambda_n$ . The germ of the optimization principles for characterizing the extreme eigenvalues is contained in this geometrical observation.

Let us turn our attention to the associated quadratic form

$$q(\mathbf{y}) = \mathbf{y}^T \Lambda \mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2. \quad (8.40)$$

<sup>†</sup> If  $\lambda_i < 0$ , then the effect is to stretch and reflect.

Note that  $q(t\mathbf{e}_1) = \lambda_1 t^2$ , and hence if  $\lambda_1 > 0$ , then  $q(\mathbf{y})$  has no maximum; on the other hand, if  $\lambda_1 \leq 0$ , so all eigenvalues are non-positive, then  $q(\mathbf{y}) \leq 0$  for all  $\mathbf{y}$ , and its maximal value is  $q(\mathbf{0}) = 0$ . Thus, in either case, a strict maximization of  $q(\mathbf{y})$  does not tell us anything of importance.

Suppose, however, that we continue in our quest to maximize  $q(\mathbf{y})$ , but now restrict  $\mathbf{y}$  to be a unit vector (in the Euclidean norm), so

$$\|\mathbf{y}\|^2 = y_1^2 + \cdots + y_n^2 = 1.$$

In view of (8.39),

$$q(\mathbf{y}) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 \leq \lambda_1 y_1^2 + \lambda_1 y_2^2 + \cdots + \lambda_1 y_n^2 = \lambda_1 (y_1^2 + \cdots + y_n^2) = \lambda_1.$$

Moreover,  $q(\mathbf{e}_1) = \lambda_1$ . We conclude that the maximal value of  $q(\mathbf{y})$  over all unit vectors is the largest eigenvalue of  $\Lambda$ :

$$\lambda_1 = \max \{ q(\mathbf{y}) \mid \|\mathbf{y}\| = 1 \}.$$

By the same reasoning, its minimal value equals the smallest eigenvalue:

$$\lambda_n = \min \{ q(\mathbf{y}) \mid \|\mathbf{y}\| = 1 \}.$$

Thus, we can characterize the two extreme eigenvalues by optimization principles, albeit of a slightly different character from what we dealt with in Chapter 5.

Now suppose  $A$  is any symmetric matrix. We use its spectral factorization (8.35) to diagonalize the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} = \mathbf{x}^T Q \Lambda Q^T \mathbf{x} = \mathbf{y}^T \Lambda \mathbf{y}, \quad \text{where} \quad \mathbf{y} = Q^T \mathbf{x} = Q^{-1} \mathbf{x},$$

as in (8.36). According to the preceding discussion, the maximum of  $\mathbf{y}^T \Lambda \mathbf{y}$  over all unit vectors  $\|\mathbf{y}\| = 1$  is the largest eigenvalue  $\lambda_1$  of  $\Lambda$ , which is the *same* as the largest eigenvalue of  $A$ . Moreover, since  $Q$  is an orthogonal matrix, Proposition 7.24 tell us that it maps unit vectors to unit vectors:

$$1 = \|\mathbf{y}\| = \|Q^T \mathbf{x}\| = \|\mathbf{x}\|,$$

and so the maximum of  $q(\mathbf{x})$  over all unit vectors  $\|\mathbf{x}\| = 1$  is the same maximum eigenvalue  $\lambda_1$ . Similar reasoning applies to the smallest eigenvalue  $\lambda_n$ . In this fashion, we have established the basic optimization principles for the extreme eigenvalues of a symmetric matrix.

**Theorem 8.40.** Let  $A$  be a symmetric matrix with real eigenvalues  $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ . Then

$$\lambda_1 = \max \{ \mathbf{x}^T A \mathbf{x} \mid \|\mathbf{x}\| = 1 \}, \quad \lambda_n = \min \{ \mathbf{x}^T A \mathbf{x} \mid \|\mathbf{x}\| = 1 \}, \quad (8.41)$$

are, respectively its largest and smallest eigenvalues. The maximal value is achieved when  $\mathbf{x} = \pm \mathbf{u}_1$  is one of the unit eigenvectors associated with the largest eigenvalue  $\lambda_1$ ; similarly, the minimal value occurs at  $\mathbf{x} = \pm \mathbf{u}_n$ , a unit eigenvector for the smallest eigenvalue  $\lambda_n$ .

**Remark.** In multivariable calculus, the eigenvalue  $\lambda$  plays the role of a *Lagrange multiplier* for the constrained optimization problem, [2, 78, 79].

**Example 8.41.** The problem is to maximize the value of the quadratic form

$$q(x, y) = 3x^2 + 2xy + 3y^2$$

for all  $x, y$  lying on the unit circle  $x^2 + y^2 = 1$ . This maximization problem is precisely of the form (8.41). The symmetric coefficient matrix for the quadratic form is  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ ,

whose eigenvalues are, according to Example 8.5,  $\lambda_1 = 4$  and  $\lambda_2 = 2$ . Theorem 8.40 implies that the maximum is the largest eigenvalue, and hence equal to 4, while its minimum is the smallest eigenvalue, and hence equal to 2. Thus, evaluating  $q(x, y)$  on the unit eigenvectors, we conclude that

$$q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = 2 \leq q(x, y) \leq 4 = q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) \quad \text{for all } x^2 + y^2 = 1.$$

In practical applications, the restriction of the quadratic form to unit vectors may not be particularly convenient. We can, however, rephrase the eigenvalue optimization principles in a form that utilizes general nonzero vectors. If  $\mathbf{v} \neq \mathbf{0}$ , then  $\mathbf{x} = \mathbf{v}/\|\mathbf{v}\|$  is a unit vector. Substituting this expression for  $\mathbf{x}$  in the quadratic form  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$  leads to the following optimization principles for the extreme eigenvalues of a symmetric matrix:

$$\lambda_1 = \max \left\{ \frac{\mathbf{v}^T A \mathbf{v}}{\|\mathbf{v}\|^2} \mid \mathbf{v} \neq \mathbf{0} \right\}, \quad \lambda_n = \min \left\{ \frac{\mathbf{v}^T A \mathbf{v}}{\|\mathbf{v}\|^2} \mid \mathbf{v} \neq \mathbf{0} \right\}. \quad (8.42)$$

Thus, we replace optimization of a quadratic polynomial over the unit sphere by optimization of a rational function over all of  $\mathbb{R}^n \setminus \{\mathbf{0}\}$ . The rational function being optimized is called a *Rayleigh quotient*, after Lord Rayleigh, a prominent nineteenth-century British scientist. For instance, referring back to Example 8.41, the maximum value of

$$r(x, y) = \frac{3x^2 + 2xy + 3y^2}{x^2 + y^2} \quad \text{for all } \begin{pmatrix} x \\ y \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is equal to 4, the same maximal eigenvalue of the corresponding coefficient matrix.

What about characterizing one of the intermediate eigenvalues? Then we need to be a little more sophisticated in designing the optimization principle. To motivate the construction, look first at the diagonal case. If we restrict the quadratic form (8.40) to vectors  $\tilde{\mathbf{y}} = (0, y_2, \dots, y_n)^T$  whose first component is zero, we obtain

$$q(\tilde{\mathbf{y}}) = q(0, y_2, \dots, y_n) = \lambda_2 y_2^2 + \dots + \lambda_n y_n^2.$$

The maximum value of  $q(\tilde{\mathbf{y}})$  over all such  $\tilde{\mathbf{y}}$  of norm 1 is, by the same reasoning, the second largest eigenvalue  $\lambda_2$ . Moreover,  $\tilde{\mathbf{y}} \cdot \mathbf{e}_1 = 0$ , and so  $\tilde{\mathbf{y}}$  can be characterized as being orthogonal to the first standard basis vector, which also happens to be the eigenvector of  $\Lambda$  corresponding to the maximal eigenvalue  $\lambda_1$ . Thus, to find the second eigenvalue, we maximize the quadratic form over all unit vectors that are orthogonal to the first eigenvector. Similarly, if we want to find the  $j^{\text{th}}$  largest eigenvalue  $\lambda_j$ , we maximize  $q(\hat{\mathbf{y}})$  over all unit vectors  $\hat{\mathbf{y}}$  whose first  $j-1$  components vanish,  $y_1 = \dots = y_{j-1} = 0$ , or, stated geometrically, over all vectors  $\hat{\mathbf{y}}$  such that  $\|\hat{\mathbf{y}}\| = 1$  and  $\hat{\mathbf{y}} \cdot \mathbf{e}_1 = \dots = \hat{\mathbf{y}} \cdot \mathbf{e}_{j-1} = 0$ , that is, over all unit vectors orthogonal to the first  $j-1$  eigenvectors of  $\Lambda$ .

A similar reasoning based on the Spectral Theorem 8.38 and the orthogonality of eigenvectors of symmetric matrices leads to the general result.

**Theorem 8.42.** Let  $A$  be a symmetric matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  and corresponding orthogonal eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_n$ . Then the maximal value of the quadratic form  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$  over all unit vectors that are orthogonal to the first  $j-1$  eigenvectors is its  $j^{\text{th}}$  eigenvalue:

$$\lambda_j = \max \left\{ \mathbf{x}^T A \mathbf{x} \mid \|\mathbf{x}\| = 1, \mathbf{x} \cdot \mathbf{v}_1 = \dots = \mathbf{x} \cdot \mathbf{v}_{j-1} = 0 \right\}. \quad (8.43)$$

Thus, at least in principle, one can compute the eigenvalues and eigenvectors of a symmetric matrix by the following recursive procedure. First, find the largest eigenvalue  $\lambda_1$  by the basic maximization principle (8.41) and its associated eigenvector  $\mathbf{v}_1$  by solving the eigenvector system (8.13). The next largest eigenvalue  $\lambda_2$  is then characterized by the constrained maximization principle (8.43), and so on. Although of theoretical interest, this algorithm is of somewhat limited value in numerical computations. Practical numerical methods for computing eigenvalues will be developed in Sections 9.5 and 9.6.

## Exercises

- 8.5.33. Find the minimum and maximum values of the quadratic form  $5x^2 + 4xy + 5y^2$  where  $x, y$  are subject to the constraint  $x^2 + y^2 = 1$ .
- 8.5.34. Find the minimum and maximum values of the quadratic form  $2x^2 + xy + 2xz + 2y^2 + 2z^2$  where  $x, y, z$  are required to satisfy  $x^2 + y^2 + z^2 = 1$ .
- 8.5.35. What is the minimum and maximum values of the following rational functions:  
 (a)  $\frac{3x^2 - 2y^2}{x^2 + y^2}$ , (b)  $\frac{x^2 - 3xy + y^2}{x^2 + y^2}$ , (c)  $\frac{3x^2 + xy + 5y^2}{x^2 + y^2}$ , (d)  $\frac{2x^2 + xy + 3xz + 2y^2 + 2z^2}{x^2 + y^2 + z^2}$ .
- 8.5.36. Find the minimum and maximum values of  $q(\mathbf{x}) = \sum_{i=1}^{n-1} x_i x_{i+1}$  for  $\|\mathbf{x}\|^2 = 1$ .  
*Hint:* See Exercise 8.2.47.
- 8.5.37. Write down and solve an optimization principle characterizing the largest and smallest eigenvalues of the following positive definite matrices:  
 (a)  $\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ , (b)  $\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$ , (c)  $\begin{pmatrix} 6 & -4 & 1 \\ -4 & 6 & -1 \\ 1 & -1 & 11 \end{pmatrix}$ , (d)  $\begin{pmatrix} 4 & -1 & -2 \\ -1 & 4 & -1 \\ -2 & -1 & 4 \end{pmatrix}$ .
- 8.5.38. Write down a maximization principle that characterizes the middle eigenvalue of the matrices in parts (c) and (d) of Exercise 8.5.37.
- 8.5.39. Given a  $3 \times 3$  symmetric matrix, formulate two distinct ways of characterizing its middle eigenvalue  $\lambda_2$ .
- 8.5.40. Suppose  $K > 0$ . What is the maximum value of  $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$  when  $\mathbf{x}$  is constrained to a sphere of radius  $\|\mathbf{x}\| = r$ ?
- 8.5.41. Let  $K > 0$ . Prove the product formula  

$$\max \left\{ \mathbf{x}^T K \mathbf{x} \mid \|\mathbf{x}\| = 1 \right\} \min \left\{ \mathbf{x}^T K^{-1} \mathbf{x} \mid \|\mathbf{x}\| = 1 \right\} = 1.$$
- ◇ 8.5.42. Write out the details in the proof of Theorem 8.42.
- ◇ 8.5.43. Reformulate Theorem 8.42 as a minimum principle for intermediate eigenvalues.
- 8.5.44. Under the set-up of Theorem 8.42, explain why  

$$\lambda_j = \max \left\{ \frac{\mathbf{v}^T K \mathbf{v}}{\|\mathbf{v}\|^2} \mid \mathbf{v} \neq \mathbf{0}, \mathbf{v} \cdot \mathbf{v}_1 = \cdots = \mathbf{v} \cdot \mathbf{v}_{j-1} = 0 \right\}.$$
- ◇ 8.5.45. (a) Let  $K, M$  be positive definite  $n \times n$  matrices and  $\lambda_1 \geq \cdots \geq \lambda_n$  be their generalized eigenvalues, as in Exercise 8.5.9. Prove that that the largest generalized eigenvalue can be characterized by the maximum principle  $\lambda_1 = \max \{ \mathbf{x}^T K \mathbf{x} \mid \mathbf{x}^T M \mathbf{x} = 1 \}$ .  
*Hint:* Use Exercise 8.5.27. (b) Prove the alternative maximum principle  $\lambda_1 = \max \left\{ \frac{\mathbf{x}^T K \mathbf{x}}{\mathbf{x}^T M \mathbf{x}} \mid \mathbf{x} \neq \mathbf{0} \right\}$ . (c) How would you characterize the smallest generalized eigenvalue? (d) An intermediate generalized eigenvalue?

8.5.46. Use Exercise 8.5.45 to find the minimum and maximum of the rational functions

$$(a) \frac{3x^2 + 2y^2}{4x^2 + 5y^2}, (b) \frac{x^2 - xy + 2y^2}{2x^2 - xy + y^2}, (c) \frac{2x^2 + 3y^2 + z^2}{x^2 + 3y^2 + 2z^2}, (d) \frac{2x^2 + 6xy + 11y^2 + 6yz + 2z^2}{x^2 + 2xy + 3y^2 + 2yz + z^2}.$$

8.5.47. Let  $A$  be a complete square matrix, not necessarily symmetric, with all positive eigenvalues. Is the associated quadratic form  $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} > 0$  for all  $\mathbf{x} \neq \mathbf{0}$ ?

## 8.6 Incomplete Matrices

Unfortunately, not all square matrices are complete. Matrices that do not have enough (complex) eigenvectors to form a basis are considerably less pleasant to work with. However, since they occasionally appear in applications, it is worth learning how to handle them. There are two approaches: the first, named after the twentieth-century Russian/German mathematician Issai Schur, is a generalization of the spectral theorem, and converts an arbitrary square matrix into a similar, upper triangular matrix with the eigenvalues along the diagonal. Thus, although not every matrix can be diagonalized, they can all be “triangularized”. Applications of the Schur decomposition, including the numerical computation of eigenvalues, can be found in [21].

The second approach, due to the nineteenth-century French mathematician Camille Jordan,<sup>†</sup> shows how to supplement the eigenvectors of an incomplete matrix in order to obtain a basis in which the matrix assumes a simple, but now non-diagonal, canonical (meaning distinguished) form. Applications of the Jordan canonical form will appear in our study of linear systems of ordinary differential equations. We remark that the two subsections are completely independent of one another.

### The Schur Decomposition

As noted above, the Schur decomposition is used to convert a square matrix into similar upper triangular matrix. The similarity transformation can be chosen to be represented by a unitary matrix — a complex generalization of an orthogonal matrix.

**Definition 8.43.** A complex, square matrix  $U$  is called *unitary* if it satisfies

$$U^\dagger U = I, \quad \text{where} \quad U^\dagger = \overline{U^T} \quad (8.44)$$

denotes the *Hermitian transpose*, in which one first transposes and then takes complex conjugates of all entries.

Thus,  $U$  is unitary if and only if its inverse equals its Hermitian transpose:  $U^{-1} = U^\dagger$ .

For example,  $U = \begin{pmatrix} \frac{1}{\sqrt{2}}i & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}i \end{pmatrix}$  is unitary, since  $U^{-1} = U^\dagger = \begin{pmatrix} -\frac{1}{\sqrt{2}}i & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}i \end{pmatrix}$ .

The  $(i, j)$  entry of the defining equation (8.44) is the Hermitian dot product between the  $i^{\text{th}}$  and  $j^{\text{th}}$  columns of  $U$ , and hence  $U$  is an  $n \times n$  unitary matrix if and only if its columns form an orthonormal basis of  $\mathbb{C}^n$ . In particular, a real matrix is unitary if and only if it is orthogonal. The next result is proved in the same fashion as Proposition 4.23.

<sup>†</sup> No relation to Wilhelm Jordan of Gauss–Jordan fame.

**Proposition 8.44.** If  $U_1$  and  $U_2$  are  $n \times n$  unitary matrices, so is their product  $U_1 U_2$ .

The *Schur decomposition* states that every square matrix is unitarily similar to an upper triangular matrix. The method of proof provides a recursive algorithm for constructing the decomposition.

**Theorem 8.45.** Let  $A$  be an  $n \times n$  matrix, either real or complex. Then there exists a unitary matrix  $U$  and an upper triangular matrix  $\Delta$  such that

$$A = U \Delta U^\dagger = U \Delta U^{-1}. \quad (8.45)$$

The diagonal entries of  $\Delta$  are the eigenvalues of  $A$ .

In particular, if all eigenvalues of  $A$  are real, then  $U = Q$  can be chosen to be a (real) orthogonal matrix, and  $\Delta$  is also a real matrix. This follows from the construction outlined in the proof.

**Warning.** The Schur decomposition (8.45) is not unique. As the method of proof makes clear, there are many inequivalent choices for the matrices  $U$  and  $\Delta$ .

*Proof of Theorem 8.45:* The proof proceeds by induction on the size of  $A$ . According to Theorem 8.11,  $A$  has at least one, possibly complex, eigenvalue  $\lambda_1$ . Let  $\mathbf{u}_1 \in \mathbb{C}^n$  be a corresponding unit eigenvector, so its Hermitian norm is  $\|\mathbf{u}_1\| = 1$ . Let  $U_1$  be an  $n \times n$  unitary matrix whose first column is the unit eigenvector  $\mathbf{u}_1$ . In practice,  $U_1$  can be constructed by applying the Gram–Schmidt process to any basis of  $\mathbb{C}^n$  whose first element is the eigenvector  $\mathbf{u}_1$ . The eigenvector equation  $A\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$  forms the first column of the matrix product equation

$$AU_1 = U_1 B, \quad \text{where} \quad B = U_1^\dagger A U_1 = \begin{pmatrix} \lambda_1 & \mathbf{r} \\ \mathbf{0} & C \end{pmatrix},$$

with  $C$  an  $(n-1) \times (n-1)$  matrix and  $\mathbf{r}$  a row vector. By our induction hypothesis, there is an  $(n-1) \times (n-1)$  unitary matrix  $V$  such that  $V^\dagger C V = \Gamma$  is an upper triangular  $(n-1) \times (n-1)$  matrix. Set  $U_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & V \end{pmatrix}$ . It is easily checked that  $U_2$  is also unitary, and, moreover,

$$U_2^\dagger B U_2 = \begin{pmatrix} \lambda_1 & \mathbf{s} \\ \mathbf{0} & \Gamma \end{pmatrix} = \Delta$$

is upper triangular. The unitary product matrix  $U = U_1 U_2$  yields the desired result:

$$U^\dagger A U = (U_1 U_2)^\dagger A U_1 U_2 = U_2^\dagger U_1^\dagger A U_1 U_2 = U_2^\dagger B U_2 = \Delta,$$

which establishes the Schur decomposition formula (8.45). Finally, since  $A$  and  $\Delta$  are similar matrices, they have the same eigenvalues, which justifies the final statement of the theorem. *Q.E.D.*

**Example 8.46.** The matrix  $A = \begin{pmatrix} 6 & 4 & -3 \\ -4 & -2 & 2 \\ 4 & 4 & -2 \end{pmatrix}$  has a simple eigenvalue  $\lambda_1 = 2$ , with

eigenvector  $\mathbf{v}_1 = (-1, 1, 0)^T$ , and a double eigenvalue  $\lambda_2 = 0$ , with only one independent eigenvector  $\mathbf{v}_2 = (1, 0, 2)^T$ . Thus  $A$  is incomplete, and so not diagonalizable. To construct a Schur decomposition, we begin with the first eigenvector  $\mathbf{v}_1$  and apply the Gram–Schmidt

process to the basis  $\mathbf{v}_1, \mathbf{e}_2, \mathbf{e}_3$  to obtain the orthogonal matrix  $U_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

The resulting similar matrix<sup>†</sup>  $B = U_1^T A U_1 = \begin{pmatrix} 2 & -8 & \frac{\sqrt{5}}{2} \\ 0 & 2 & -\frac{1}{\sqrt{2}} \\ 0 & 4\sqrt{2} & -2 \end{pmatrix}$  has its first column in

upper triangular form. To continue the procedure, we extract the lower  $2 \times 2$  submatrix  $C = \begin{pmatrix} 2 & -\frac{1}{\sqrt{2}} \\ 4\sqrt{2} & -2 \end{pmatrix}$ , and find that it has a single (incomplete) eigenvalue 0, with unit eigenvector  $\begin{pmatrix} \frac{1}{3} \\ \frac{2\sqrt{2}}{3} \\ \frac{2}{3} \end{pmatrix}$ . The corresponding orthogonal matrix  $V = \begin{pmatrix} \frac{1}{3} & \frac{2\sqrt{2}}{3} \\ \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{pmatrix}$  will con-

vert  $C$  to upper triangular form  $V^T C V = \begin{pmatrix} 0 & \frac{9}{\sqrt{2}} \\ 0 & 0 \end{pmatrix}$ . Therefore,  $U_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{2\sqrt{2}}{3} \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{pmatrix}$

will complete the conversion of the original matrix into upper triangular form

$\Delta = U_2^T B U_2 = U^T A U = \begin{pmatrix} 2 & \frac{2}{3} & -\frac{37}{3\sqrt{2}} \\ 0 & 0 & \frac{9}{\sqrt{2}} \\ 0 & 0 & 0 \end{pmatrix}$ , where  $U = U_1 U_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{3} \\ \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & -\frac{1}{3} \end{pmatrix}$

is the desired orthogonal (unitary) matrix. Use of a computer to carry out the detailed calculations is essential in most examples.

## Exercises

- 8.6.1. Establish a Schur Decomposition for the matrices (a)  $\begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$ ,  
 (c)  $\begin{pmatrix} 8 & 9 \\ -6 & -7 \end{pmatrix}$ , (d)  $\begin{pmatrix} 1 & 5 \\ -2 & -1 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & -1 & 2 \\ -2 & 3 & -1 \\ -6 & 6 & -5 \end{pmatrix}$ , (f)  $\begin{pmatrix} 0 & 2 & -1 \\ -1 & -1 & 1 \\ -1 & 0 & 0 \end{pmatrix}$ .

- 8.6.2. Show that a real unitary matrix is an orthogonal matrix.

- ◇ 8.6.3. Prove Proposition 8.44.

- ◇ 8.6.4. Write out a new proof of the Spectral Theorem 8.38 based on the Schur Decomposition.

- ◇ 8.6.5. A complex matrix  $A$  is called *normal* if it commutes with its Hermitian transpose:

$A^\dagger A = A A^\dagger$ . (a) Show that every real symmetric matrix is normal. (b) Show that every unitary matrix is normal. (c) Show that every real orthogonal matrix is normal. (d) Show that an upper triangular matrix is normal if and only if it is diagonal. (e) Show that the eigenvectors of a normal matrix form an orthogonal basis of  $\mathbb{C}^n$  under the Hermitian dot product. (f) Show that the converse is true: a matrix has an orthogonal eigenvector basis of  $\mathbb{C}^n$  if and only if it is normal. *Hint:* Use the Schur Decomposition. (g) How can you tell when a real matrix has a real orthonormal eigenvector basis?

<sup>†</sup> Since all matrices are real in this example, the Hermitian transpose  ${}^\dagger$  reduces to the ordinary transpose  ${}^T$ .

## The Jordan Canonical Form

We now turn to the more sophisticated Jordan canonical form. Throughout this section,  $A$  will be an  $n \times n$  matrix, with either real or complex entries. We let  $\lambda_1, \dots, \lambda_k$  denote the *distinct* eigenvalues of  $A$ , so  $\lambda_i \neq \lambda_j$  for  $i \neq j$ . We recall that Theorem 8.11 guarantees that every matrix has at least one (complex) eigenvalue, so  $k \geq 1$ . Moreover, we are assuming that  $k < n$ , since otherwise  $A$  would be complete.

**Definition 8.47.** Let  $A$  be a square matrix. A *Jordan chain* of length  $j$  for  $A$  is a sequence of non-zero vectors  $\mathbf{w}_1, \dots, \mathbf{w}_j \in \mathbb{C}^n$  that satisfies

$$A\mathbf{w}_1 = \lambda\mathbf{w}_1, \quad A\mathbf{w}_i = \lambda\mathbf{w}_i + \mathbf{w}_{i-1}, \quad i = 2, \dots, j, \quad (8.46)$$

where  $\lambda$  is an eigenvalue of  $A$ . A Jordan chain associated with a zero eigenvalue, which requires that  $A$  be singular, is called a *null Jordan chain*, and satisfies

$$A\mathbf{w}_1 = \mathbf{0}, \quad A\mathbf{w}_i = \mathbf{w}_{i-1}, \quad i = 2, \dots, j. \quad (8.47)$$

Note that the initial vector  $\mathbf{w}_1$  in a Jordan chain is a genuine eigenvector, and so Jordan chains exist only when  $\lambda$  is an eigenvalue. The rest,  $\mathbf{w}_2, \dots, \mathbf{w}_j$ , are *generalized eigenvectors*, in accordance with the following definition.

**Definition 8.48.** A nonzero vector  $\mathbf{w} \neq \mathbf{0}$  such that

$$(A - \lambda \mathbf{I})^k \mathbf{w} = \mathbf{0} \quad (8.48)$$

for some  $k > 0$  and  $\lambda \in \mathbb{C}$  is called a *generalized eigenvector*<sup>†</sup> of the matrix  $A$ .

Note that every ordinary eigenvector is automatically a generalized eigenvector, since we can just take  $k = 1$  in (8.48); but the converse is not necessarily valid. We shall call the minimal value of  $k$  for which (8.48) holds the *index* of the generalized eigenvector. Thus, an ordinary eigenvector is a generalized eigenvector of index 1. Since  $A - \lambda \mathbf{I}$  is nonsingular whenever  $\lambda$  is not an eigenvalue of  $A$ , its  $k^{\text{th}}$  power  $(A - \lambda \mathbf{I})^k$  is also nonsingular. Therefore, generalized eigenvectors can exist only when  $\lambda$  is an ordinary eigenvalue of  $A$  — there are no additional “generalized eigenvalues”.

**Lemma 8.49.** The  $i^{\text{th}}$  vector  $\mathbf{w}_i$  in a Jordan chain (8.46) is a generalized eigenvector of index  $i$ .

*Proof:* By definition,  $(A - \lambda \mathbf{I})\mathbf{w}_1 = \mathbf{0}$ , and so  $\mathbf{w}_1$  is an eigenvector. Next, we have  $(A - \lambda \mathbf{I})\mathbf{w}_2 = \mathbf{w}_1 \neq \mathbf{0}$ , while  $(A - \lambda \mathbf{I})^2\mathbf{w}_2 = (A - \lambda \mathbf{I})\mathbf{w}_1 = \mathbf{0}$ . Thus,  $\mathbf{w}_2$  is a generalized eigenvector of index 2. A simple induction proves that  $(A - \lambda \mathbf{I})^{i-1}\mathbf{w}_i = \mathbf{w}_1 \neq \mathbf{0}$  while  $(A - \lambda \mathbf{I})^i\mathbf{w}_i = \mathbf{0}$ . *Q.E.D.*

**Example 8.50.** Consider the  $3 \times 3$  matrix  $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$ . The only eigenvalue is  $\lambda = 2$ , and  $A - 2\mathbf{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ . We claim that the standard basis vectors  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$

<sup>†</sup> Despite the common terminology, this is *not* the same concept as developed in Exercise 8.5.8.

form a Jordan chain. Indeed,  $A\mathbf{e}_1 = 2\mathbf{e}_1$ , and hence  $\mathbf{e}_1 \in \ker(A - 2\mathbf{I})$  is a genuine eigenvector. Furthermore,  $A\mathbf{e}_2 = 2\mathbf{e}_2 + \mathbf{e}_1$ , and  $A\mathbf{e}_3 = 2\mathbf{e}_3 + \mathbf{e}_2$ , as you can easily check. Thus,  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  satisfy the Jordan chain equations (8.46) for the eigenvalue  $\lambda = 2$ . Note

that  $\mathbf{e}_2$  lies in the kernel of  $(A - 2\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ , and so is a generalized eigenvector

of index 2. Indeed, every vector of the form  $\mathbf{w} = a\mathbf{e}_1 + b\mathbf{e}_2$  with  $b \neq 0$  is a generalized eigenvector of index 2. (When  $b = 0, a \neq 0$ , the vector  $\mathbf{w} = a\mathbf{e}_1$  is an ordinary eigenvector of index 1.) Finally,  $(A - 2\mathbf{I})^3 = \mathbf{0}$ , and so every nonzero vector  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^3$ , including  $\mathbf{e}_3$ , is a generalized eigenvector of index 3 or less.

A basis of  $\mathbb{R}^n$  or  $\mathbb{C}^n$  is called a *Jordan basis* for the matrix  $A$  if it consists of one or more Jordan chains that have no elements in common. Thus, for the matrix in Example 8.50, the standard basis  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  is, in fact, a Jordan basis. An eigenvector basis, if such exists, qualifies as a Jordan basis, since each eigenvector belongs to a Jordan chain of length 1. Jordan bases are the desired extension of eigenvector bases, and every square matrix has one. The proof of the following *Jordan Basis Theorem* will appear below.

**Theorem 8.51.** Every  $n \times n$  matrix admits a Jordan basis of  $\mathbb{C}^n$ . The first elements of the Jordan chains form a maximal set of linearly independent eigenvectors. Moreover, the number of generalized eigenvectors in the Jordan basis that belong to the Jordan chains associated with the eigenvalue  $\lambda$  is the same as the eigenvalue's multiplicity.

**Example 8.52.** Consider the  $5 \times 5$  matrix  $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$ . Its characteristic equation is found to be

$$p_A(\lambda) = \det(A - \lambda\mathbf{I}) = \lambda^5 + \lambda^4 - 5\lambda^3 - \lambda^2 + 8\lambda - 4 = (\lambda - 1)^3(\lambda + 2)^2 = 0,$$

and hence  $A$  has two eigenvalues:  $\lambda_1 = 1$ , which is a triple eigenvalue, and  $\lambda_2 = -2$ , which is double. Solving the associated homogeneous systems  $(A - \lambda_j\mathbf{I})\mathbf{v} = \mathbf{0}$ , we find that, up to constant multiple, there are only two eigenvectors:  $\mathbf{v}_1 = (0, 0, 0, -1, 1)^T$  for  $\lambda_1 = 1$  and, anticipating our final numbering,  $\mathbf{v}_4 = (-1, 1, 1, -2, 0)^T$  for  $\lambda_2 = -2$ . Thus,  $A$  is far from complete.

To construct a Jordan basis, we first note that since  $A$  has 2 linearly independent eigenvectors, the Jordan basis will contain two Jordan chains: the one associated with the triple eigenvalue  $\lambda_1 = 1$  will have length 3, while  $\lambda_2 = -2$  leads to a Jordan chain of length 2. To construct the former, we need to first solve the system  $(A - \mathbf{I})\mathbf{w} = \mathbf{v}_1$ . Note that the coefficient matrix is singular — it must be, since 1 is an eigenvalue — and the general solution is  $\mathbf{w} = \mathbf{v}_2 + t\mathbf{v}_1$ , where  $\mathbf{v}_2 = (0, 1, 0, 0, -1)^T$ , and  $t$  is the free variable. The appearance of an arbitrary multiple of the eigenvector  $\mathbf{v}_1$  in the solution is not unexpected; indeed, the kernel of  $A - \mathbf{I}$  is the eigenspace for  $\lambda_1 = 1$ . We can choose any solution, e.g.,  $\mathbf{v}_2$  as the second element in the Jordan chain. We then solve  $(A - \mathbf{I})\mathbf{w} = \mathbf{v}_2$  for  $\mathbf{w} = \mathbf{v}_3 + t\mathbf{v}_1$ , where  $\mathbf{v}_3 = (0, 0, 0, 1, 0)^T$  can be used as the final element of this Jordan chain. Similarly, to construct the Jordan chain for the second eigenvalue, we solve  $(A + 2\mathbf{I})\mathbf{w} = \mathbf{v}_4$  for  $\mathbf{w} = \mathbf{v}_5 + t\mathbf{v}_4$ , where  $\mathbf{v}_5 = (-1, 0, 0, -2, 1)^T$ .

Thus, the desired Jordan basis is

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -2 \\ 0 \end{pmatrix}, \quad \mathbf{v}_5 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ -2 \\ 1 \end{pmatrix},$$

with  $A\mathbf{v}_1 = \mathbf{v}_1$ ,  $A\mathbf{v}_2 = \mathbf{v}_2 + \mathbf{v}_1$ ,  $A\mathbf{v}_3 = \mathbf{v}_3 + \mathbf{v}_2$ ,  $A\mathbf{v}_4 = -2\mathbf{v}_4$ ,  $A\mathbf{v}_5 = -2\mathbf{v}_5 + \mathbf{v}_4$ .

Just as an eigenvector basis diagonalizes a complete matrix, a Jordan basis provides a particularly simple form for an incomplete matrix, known as the *Jordan canonical form*.

**Definition 8.53.** An  $n \times n$  matrix of the form<sup>†</sup>

$$J_{\lambda,n} = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & 1 & \\ & & & \ddots & \ddots \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \quad (8.49)$$

in which  $\lambda$  is a real or complex number, is known as a *Jordan block*.

In particular, a  $1 \times 1$  Jordan block is merely a scalar  $J_{\lambda,1} = \lambda$ . Since every matrix has at least one (complex) eigenvector — see Theorem 8.11 — the Jordan block matrices have the least possible number of independent eigenvectors. The following result is immediate.

**Lemma 8.54.** The  $n \times n$  Jordan block matrix  $J_{\lambda,n}$  has a single eigenvalue,  $\lambda$ , and a single independent eigenvector,  $\mathbf{e}_1$ . The standard basis vectors  $\mathbf{e}_1, \dots, \mathbf{e}_n$  form a Jordan chain for  $J_{\lambda,n}$ .

**Definition 8.55.** A *Jordan matrix* is a square matrix of block diagonal form

$$J = \text{diag}(J_{\lambda_1, n_1}, J_{\lambda_2, n_2}, \dots, J_{\lambda_k, n_k}) = \begin{pmatrix} J_{\lambda_1, n_1} & & & \\ & J_{\lambda_2, n_2} & & \\ & & \ddots & \\ & & & J_{\lambda_k, n_k} \end{pmatrix}, \quad (8.50)$$

in which one or more Jordan blocks, not necessarily of the same size, lie along the diagonal, while the blank off-diagonal blocks are all zero.

Note that the only possibly non-zero entries in a Jordan matrix are those on the diagonal, which can have any complex value, including 0, and those on the superdiagonal, which are either 1 or 0. The positions of the superdiagonal 1's uniquely prescribes the Jordan blocks.

---

<sup>†</sup> All non-displayed entries are zero.

For example, the  $6 \times 6$  matrices

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

are all Jordan matrices: the first is a diagonal matrix, consisting of 6 distinct  $1 \times 1$  Jordan blocks; the second has a  $4 \times 4$  Jordan block followed by a  $2 \times 2$  block that happen to have the same diagonal entries; the last has three  $2 \times 2$  Jordan blocks with respective diagonal entries 0, 1, 2.

As a direct corollary of Lemma 8.54 combined with the matrix's block structure, cf. Exercise 8.2.50, we obtain a complete classification of the eigenvectors and eigenvalues of a Jordan matrix.

**Lemma 8.56.** The Jordan matrix (8.50) has eigenvalues  $\lambda_1, \dots, \lambda_k$ . The standard basis vectors  $\mathbf{e}_1, \dots, \mathbf{e}_n$  form a Jordan basis, which is partitioned into nonoverlapping Jordan chains labeled by the Jordan blocks.

Thus, in the preceding examples of Jordan matrices, the first has three double eigenvalues, 1 and 2 and 3, and corresponding linearly independent eigenvectors  $\mathbf{e}_1, \mathbf{e}_6$ , and  $\mathbf{e}_2, \mathbf{e}_5$ , and  $\mathbf{e}_3, \mathbf{e}_4$ , each of which belongs to a Jordan chain of length 1. The second matrix has only one eigenvalue,  $-1$ , but two independent eigenvectors  $\mathbf{e}_1, \mathbf{e}_5$ , and hence two Jordan chains, namely  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ , and  $\mathbf{e}_5, \mathbf{e}_6$ . The last has three eigenvalues 0, 1, 2, three eigenvectors  $\mathbf{e}_1, \mathbf{e}_3, \mathbf{e}_5$ , and three Jordan chains of length 2:  $\mathbf{e}_1, \mathbf{e}_2$ , and  $\mathbf{e}_3, \mathbf{e}_4$ , and  $\mathbf{e}_5, \mathbf{e}_6$ . In particular, the only complete Jordan matrices are the diagonal matrices, all of whose Jordan blocks are of size  $1 \times 1$ .

The Jordan canonical form follows from the Jordan Basis Theorem 8.51.

**Theorem 8.57.** Let  $A$  be an  $n \times n$  real or complex matrix. Let  $S = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n)$  be a matrix whose columns form a Jordan basis of  $A$ . Then  $S$  places  $A$  into the *Jordan canonical form*

$$S^{-1}AS = J = \text{diag}(J_{\lambda_1, n_1}, J_{\lambda_2, n_2}, \dots, J_{\lambda_k, n_k}), \quad \text{or, equivalently, } A = SJS^{-1}. \quad (8.51)$$

The diagonal entries of the similar Jordan matrix  $J$  are the eigenvalues of  $A$ . In particular,  $A$  is complete (diagonalizable) if and only if every Jordan block is of size  $1 \times 1$  or, equivalently, all Jordan chains are of length 1. The Jordan canonical form of  $A$  is uniquely determined up to a permutation of the diagonal Jordan blocks.

For instance, the matrix  $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$  considered in Example 8.52

has the following Jordan basis matrix and Jordan canonical form

$$S = \begin{pmatrix} 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & -2 & -2 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}, \quad J = S^{-1}AS = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}.$$

Finally, let us outline a proof of the Jordan Basis Theorem 8.51.

**Lemma 8.58.** If  $\mathbf{v}_1, \dots, \mathbf{v}_n$  forms a Jordan basis for the matrix  $A$ , it also forms a Jordan basis for  $B = A - c\mathbf{I}$ , for any scalar  $c$ .

*Proof:* We note that the eigenvalues of  $B$  are all of the form  $\lambda - c$ , where  $\lambda$  is an eigenvalue of  $A$ . Moreover, given a Jordan chain  $\mathbf{w}_1, \dots, \mathbf{w}_j$  of  $A$ , we have

$$B\mathbf{w}_1 = (\lambda - c)\mathbf{w}_1, \quad B\mathbf{w}_i = (\lambda - c)\mathbf{w}_i + \mathbf{w}_{i-1}, \quad i = 2, \dots, j,$$

so  $\mathbf{w}_1, \dots, \mathbf{w}_j$  is also a Jordan chain for  $B$  corresponding to the eigenvalue  $\lambda - c$ . *Q.E.D.*

The proof of Theorem 8.51 will be done by induction on the size  $n$  of the matrix. The case  $n = 1$  is trivial, since every nonzero element of  $\mathbb{C}$  is a Jordan basis for a  $1 \times 1$  matrix. To perform the induction step, we assume that the result is valid for all matrices of size  $\leq n-1$ . Let  $A$  be an  $n \times n$  matrix. According to Theorem 8.11,  $A$  has at least one complex eigenvalue  $\lambda_1$ . Let  $B = A - \lambda_1\mathbf{I}$ . Since  $\lambda_1$  is an eigenvalue of  $A$ , we know that 0 is an eigenvalue of  $B$ . This means that  $\ker B \neq \{\mathbf{0}\}$ , and so  $r = \text{rank } B < n$ . Moreover, by Lemma 8.58, a Jordan basis of  $B$  is also a Jordan basis for  $A$ , and so we can concentrate all our attention on the singular matrix  $B$  from now on.

Recall that  $W = \text{img } B \subset \mathbb{C}^n$  is an invariant subspace for  $B$ , i.e.,  $B\mathbf{w} \in W$  whenever  $\mathbf{w} \in W$ . Moreover, since  $B$  is singular,  $\dim W = r = \text{rank } B < n$ . Thus, by fixing a basis of  $W$ , we can realize the restriction  $B: W \rightarrow W$  as multiplication by an  $r \times r$  matrix. The fact that  $r < n$  allows us to invoke the induction hypothesis and deduce the existence of a Jordan basis  $\mathbf{w}_1, \dots, \mathbf{w}_r \in W \subset \mathbb{C}^n$  for the action of  $B$  on the subspace  $W$ . Our goal is to complete this collection to a full Jordan basis on  $\mathbb{C}^n$ .

To this end, we append two additional kinds of vectors. Suppose that the Jordan basis of  $W$  contains  $k$  null Jordan chains associated with its zero eigenvalue. Each null Jordan chain consists of vectors  $\mathbf{w}_1, \dots, \mathbf{w}_j \in W$  satisfying  $B\mathbf{w}_1 = \mathbf{0}, B\mathbf{w}_2 = \mathbf{w}_1, \dots, B\mathbf{w}_j = \mathbf{w}_{j-1}$ , cf. (8.47). The number  $k$  of null Jordan chains is equal to the number of linearly independent null eigenvectors of  $B$  that belong to  $W = \text{img } B$ , that is  $k = \dim(\ker B \cap \text{img } B)$ . To each such null Jordan chain, we append a vector  $\mathbf{w}_{j+1} \in \mathbb{C}^n$  such that  $B\mathbf{w}_{j+1} = \mathbf{w}_j$ , noting that  $\mathbf{w}_{j+1}$  exists because  $\mathbf{w}_j \in \text{img } B$ . We deduce that  $\mathbf{w}_1, \dots, \mathbf{w}_{j+1} \in \mathbb{C}^n$  forms a null Jordan chain, of length  $j+1$ . Having extended all the null Jordan chains in  $W$ , the resulting collection consists of  $r+k$  vectors in  $\mathbb{C}^n$  arranged in nonoverlapping Jordan chains. To complete to a basis, we append  $n-r-k$  additional linearly independent null vectors  $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k} \in \ker B \setminus \text{img } B$  that lie outside its image. Since  $B\mathbf{z}_j = \mathbf{0}$ , each  $\mathbf{z}_j$  forms a null Jordan chain of length 1. We claim that the complete collection consisting of the  $r$  non-null Jordan chains in  $W$ , the  $k$  extended null chains, and the additional null vectors  $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k}$ , forms the desired Jordan basis. By construction, it consists of nonoverlapping Jordan chains. The only remaining issue is the proof that the resulting collection of vectors is linearly independent; this is left as a challenge for the reader in Exercise 8.6.25. *Q.E.D.*

With the Jordan canonical form in hand, the general result characterizing complex invariant subspaces of a general square matrix can now be stated.

**Theorem 8.59.** Every complex invariant subspace of a square matrix  $A$  is spanned by a finite number of Jordan chains.

*Proof:* The proof proceeds in the same manner as that of Theorem 8.30. We assume that  $\mathbf{v}_k$  is the *last* vector in its Jordan chain. We deduce that  $A\mathbf{w} - \lambda_k\mathbf{w}$  is a linear combination of  $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$  with non-vanishing coefficients, and so the induction step remains valid in this case. The details are left to the reader. *Q.E.D.*

Similarly, a complex conjugate pair of Jordan chains of order  $k$  produces two distinct  $k$ -dimensional complex invariant subspaces, and hence a real invariant subspace of dimension  $2k$  obtained from their real and imaginary parts. A detailed statement of the nature of the most general real invariant subspace of a matrix is also left to the reader.

## Exercises

8.6.6. For each of the following Jordan matrices, identify the Jordan blocks. Write down the eigenvalues, the eigenvectors, and the Jordan basis. Clearly identify the Jordan chains.

$$(a) \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, (b) \begin{pmatrix} -3 & 0 \\ 0 & 6 \end{pmatrix}, (c) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, (d) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, (e) \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

8.6.7. Find a Jordan basis and Jordan canonical form for each of the following matrices:

$$(a) \begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}, (b) \begin{pmatrix} -1 & -1 \\ 4 & -5 \end{pmatrix}, (c) \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

$$(d) \begin{pmatrix} -3 & 1 & 0 \\ 1 & -3 & -1 \\ 0 & 1 & -3 \end{pmatrix}, (e) \begin{pmatrix} -1 & 1 & 1 \\ -2 & -2 & -2 \\ 1 & -1 & -1 \end{pmatrix}, (f) \begin{pmatrix} 2 & -1 & 1 & 2 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

8.6.8. Write down all possible  $4 \times 4$  Jordan matrices that have only  $\lambda = 2$  as an eigenvalue.

8.6.9. Write down all  $3 \times 3$  Jordan matrices that have eigenvalues 2 and 5 (and no others).

8.6.10. Write down a formula for the inverse of a Jordan block matrix.

*Hint:* Try some small examples first to help in figuring out the pattern.

8.6.11. *True or false:* If  $A$  is complete, every generalized eigenvector is an ordinary eigenvector.

8.6.12. *True or false:* Every generalized eigenvector belongs to a Jordan chain.

8.6.13. *True or false:* If  $\mathbf{w}$  is a generalized eigenvector of  $A$ , then  $\mathbf{w}$  is a generalized eigenvector of every power  $A^j$ , for  $j \in \mathbb{N}$ , thereof.

8.6.14. *True or false:* If  $\mathbf{w}$  is a generalized eigenvector of index  $k$  of  $A$ , then  $\mathbf{w}$  is an ordinary eigenvector of  $A^k$ .

◇ 8.6.15. Suppose you know all eigenvalues of a matrix as well as their algebraic and geometric multiplicities. Can you determine the matrix's Jordan canonical form?

8.6.16. *True or false:* If  $\mathbf{w}_1, \dots, \mathbf{w}_j$  is a Jordan chain for a matrix  $A$ , so are the scalar multiples  $c\mathbf{w}_1, \dots, c\mathbf{w}_j$  for all  $c \neq 0$ .

8.6.17. Let  $A$  and  $B$  be  $n \times n$  matrices. According to Exercise 8.2.23, the matrix products  $AB$  and  $BA$  have the same eigenvalues. Do they have the same Jordan form?

8.6.18. *True or false:* If the Jordan canonical form of  $A$  is  $J$ , then that of  $A^2$  is  $J^2$ .

◇ 8.6.19. (a) Give an example of a matrix  $A$  such that  $A^2$  has an eigenvector that is *not* an eigenvector of  $A$ . (b) Show that, in general, every eigenvalue of  $A^2$  is the square of an eigenvalue of  $A$ .

◇ 8.6.20. (a) Prove that a Jordan block matrix  $J_{0,n}$  with zero diagonal entries is nilpotent, as in Exercise 1.3.12. (b) Prove that a Jordan matrix is nilpotent if and only if all its diagonal entries are zero. (c) Prove that a matrix is nilpotent if and only if its Jordan canonical form is nilpotent. (d) Explain why a matrix is nilpotent if and only if its only eigenvalue is 0.

8.6.21. Let  $J$  be a Jordan matrix. (a) Prove that  $J^k$  is a complete matrix for some  $k \geq 1$  if and only if either  $J$  is diagonal, or  $J$  is nilpotent with  $J^k = O$ . (b) Suppose that  $A$  is an incomplete matrix such that  $A^k$  is complete for some  $k \geq 2$ . Prove that  $A^k = O$ , and hence  $A$  is nilpotent. (A simpler version of this problem appears in Exercise 8.3.8.)

◇ 8.6.22. *Cayley–Hamilton Theorem:* Let  $p_A(\lambda) = \det(A - \lambda I)$  be the characteristic polynomial of  $A$ . (a) Prove that if  $D$  is a diagonal matrix, then<sup>†</sup>  $p_D(D) = O$ .

*Hint:* Leave  $p_D(\lambda)$  in factored form. (b) Prove that if  $A$  is complete, then  $p_A(A) = O$ .

(c) Prove that if  $J$  is a Jordan block, then  $p_J(J) = O$ . (d) Prove that this also holds if  $J$  is a Jordan matrix. (e) Prove the Cayley–Hamilton Theorem: a square matrix  $A$  satisfies its own characteristic equation:  $p_A(A) = O$ .

◇ 8.6.23. *Minimal polynomial:* Let  $A$  be an  $n \times n$  matrix. By definition, the *minimal polynomial* of  $A$  is the monic polynomial  $m_A(t) = t^k + c_{k-1}t^{k-1} + \cdots + c_1t + c_0$  of *minimal degree*  $k$  that annihilates  $A$ , so  $m_A(A) = A^k + c_{k-1}A^{k-1} + \cdots + c_1A + c_0I = O$ .  
 (a) Prove that the monic minimal polynomial  $m_A$  is unique. (b) (c) Prove that if  $r(t)$  is any other polynomial such that  $r(A) = O$ , then  $r(t) = q(t)m_A(t)$  for some polynomial  $q(t)$ .  
 (d) Prove that the matrix's minimal polynomial is a factor of its characteristic polynomial, so  $p_A(t) = q_A(t)m_A(t)$  for some polynomial  $q_A(t)$ . *Hint:* Use the Cayley–Hamilton Theorem in Exercise 8.6.22. (e) Prove that if  $A$  has all distinct eigenvalues, then  $p_A = m_A$ .  
 (f) Prove that  $p_A = m_A$  if and only if no two Jordan blocks have the same eigenvalue.

◇ 8.6.24. Prove Lemma 8.54.

◇ 8.6.25. Prove that the  $n$  vectors constructed in the proof of Theorem 8.51 are linearly independent and hence form a Jordan basis.

*Hint:* Suppose that some linear combination vanishes. Apply  $B$  to the equation, and then use the fact that we started with a Jordan basis for  $W = \text{img } B$ .

8.6.26. Find all invariant subspaces of the following matrices: (a)  $\begin{pmatrix} 4 & 0 \\ -3 & 4 \end{pmatrix}$ , (b)  $\begin{pmatrix} -1 & -1 \\ 4 & -5 \end{pmatrix}$ ,  
 (c)  $\begin{pmatrix} 0 & -1 & 1 \\ 1 & 1 & -1 \\ 3 & 3 & -4 \end{pmatrix}$ , (d)  $\begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ , (e)  $\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} -1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ -1 & -4 & 1 & -2 \\ 0 & 1 & 0 & 1 \end{pmatrix}$ .

◇ 8.6.27. The method for constructing a Jordan basis in Example 8.52 is simplified due to the fact that each eigenvalue admits only one Jordan block. On the other hand, the method used in the proof of Theorem 8.51 is rather impractical. Devise a general method for constructing a Jordan basis for an arbitrary matrix, paying careful attention to how the Jordan chains having the same eigenvalue are found.

---

<sup>†</sup> See Exercise 1.2.35 for the basics of matrix polynomials.

## 8.7 Singular Values

We have already expounded on the central role played by the eigenvalues and eigenvectors of a square matrix in both theory and applications. Further evidence will appear in the subsequent chapters. Alas, rectangular matrices do not have eigenvalues (why?), and so, at first glance, do not appear to possess any quantities of comparable significance. But you no doubt recall that our earlier treatment of least squares minimization problems, as well as the equilibrium equations for structures and circuits, made essential use of the symmetric, positive semi-definite *square* Gram matrix  $K = A^T A$  — which can be naturally formed even when  $A$  is not square. Perhaps the eigenvalues of  $K$  might play a comparably important role for general matrices. Since they are not easily related to the eigenvalues of  $A$  — which, in the non-square case, don't even exist — we shall endow them with a new name. They were first studied by the German mathematician Erhard Schmidt in early days of the twentieth century, although intimations can be found in Gauss's work on rigid body dynamics.

**Definition 8.60.** The *singular values*  $\sigma_1, \dots, \sigma_r$  of an  $m \times n$  matrix  $A$  are the positive square roots,  $\sigma_i = \sqrt{\lambda_i} > 0$ , of the nonzero eigenvalues of the associated Gram matrix  $K = A^T A$ . The corresponding eigenvectors of  $K$  are known as the *singular vectors* of  $A$ .

Since  $K$  is necessarily positive semi-definite, its eigenvalues are always non-negative,  $\lambda_i \geq 0$ , which justifies the positivity of the singular values of  $A$  — independently of whether  $A$  itself has positive, negative, or even complex eigenvalues, or is rectangular and has no eigenvalues at all. We will follow the standard convention, and always label the singular values in decreasing order, so that  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . Thus,  $\sigma_1$  will always denote the largest, or *dominant*, singular value. If  $K = A^T A$  has repeated eigenvalues, the singular values of  $A$  are repeated with the same multiplicities. As we will see, the number  $r$  of singular values is equal to the rank of the matrices  $A$  and  $K$ .

**Warning.** Some texts include the zero eigenvalues of  $K$  as singular values of  $A$ . We find this to be less convenient for our development, but you should be aware of the differences between the two conventions.

**Example 8.61.** Let  $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$ . The associated Gram matrix

$$K = A^T A = \begin{pmatrix} 3 & 4 \\ 5 & 0 \end{pmatrix} \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$$

has eigenvalues  $\lambda_1 = 40$ ,  $\lambda_2 = 10$ , and corresponding eigenvectors  $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ .

Thus, the singular values of  $A$  are  $\sigma_1 = \sqrt{40} \approx 6.3246$  and  $\sigma_2 = \sqrt{10} \approx 3.1623$ , with  $\mathbf{v}_1, \mathbf{v}_2$  being the singular vectors. Note that the singular values are *not* its eigenvalues, which are  $\lambda_1 = \frac{1}{2}(3 + \sqrt{89}) \simeq 6.2170$  and  $\lambda_2 = \frac{1}{2}(3 - \sqrt{89}) \simeq -3.2170$ , nor are the singular vectors eigenvectors of  $A$ .

Only in the special case of symmetric matrices is there a direct connection between their singular values and their (necessarily real) eigenvalues.

**Proposition 8.62.** If  $A = A^T$  is a symmetric matrix, its singular values are the absolute values of its nonzero eigenvalues:  $\sigma_i = |\lambda_i| > 0$ ; its singular vectors coincide with its non-null eigenvectors.

*Proof:* When  $A$  is symmetric,  $K = A^T A = A^2$ . So, if

$$A\mathbf{v} = \lambda\mathbf{v}, \quad \text{then} \quad K\mathbf{v} = A^2\mathbf{v} = A(\lambda\mathbf{v}) = \lambda A\mathbf{v} = \lambda^2\mathbf{v}.$$

Thus, every eigenvector  $\mathbf{v}$  of  $A$  is also an eigenvector of  $K$  with eigenvalue  $\lambda^2$ . Therefore, the eigenvector basis of  $A$  guaranteed by Theorem 8.32 is also an eigenvector basis for  $K$ , and hence forms a complete system of singular vectors for  $A$ .  $Q.E.D.$

The generalization of the spectral factorization (8.35) to non-symmetric matrices is known as the *singular value decomposition*, commonly abbreviated SVD. Unlike the former, which applies only to square matrices, every nonzero matrix possesses a singular value decomposition.

**Theorem 8.63.** A nonzero real  $m \times n$  matrix  $A$  of rank  $r > 0$  can be factored,

$$A = P\Sigma Q^T, \quad (8.52)$$

into the product of an  $m \times r$  matrix  $P$  with orthonormal<sup>†</sup> columns, so  $P^T P = I$ , the  $r \times r$  diagonal matrix  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$  that has the singular values of  $A$  as its diagonal entries, and an  $r \times n$  matrix  $Q^T$  with orthonormal rows, so  $Q^T Q = I$ .

*Proof:* Let's begin by rewriting the desired factorization (8.52) as  $AQ = P\Sigma$ . The individual columns of this matrix equation are the vector equations

$$A\mathbf{q}_i = \sigma_i \mathbf{p}_i, \quad i = 1, \dots, r, \quad (8.53)$$

relating the orthonormal columns of  $Q = (\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_r)$  to the orthonormal columns of  $P = (\mathbf{p}_1 \ \mathbf{p}_2 \ \dots \ \mathbf{p}_r)$ . Thus, our goal is to find vectors  $\mathbf{p}_1, \dots, \mathbf{p}_r$  and  $\mathbf{q}_1, \dots, \mathbf{q}_r$  that satisfy (8.53). To this end, we let  $\mathbf{q}_1, \dots, \mathbf{q}_r$  be orthonormal eigenvectors of the Gram matrix  $K = A^T A$  corresponding to the non-zero eigenvalues, which, according to Proposition 8.37 form a basis for  $\text{img } K = \text{coimg } A$ , of dimension  $r = \text{rank } K = \text{rank } A$ . Thus, by the definition of the singular values,

$$A^T A \mathbf{q}_i = K \mathbf{q}_i = \sigma_i^2 \mathbf{q}_i, \quad i = 1, \dots, r. \quad (8.54)$$

We claim that the image vectors  $\mathbf{w}_i = A\mathbf{q}_i$  are automatically orthogonal. Indeed, in view of the orthonormality of the  $\mathbf{q}_i$  combined with (8.54),

$$\mathbf{w}_i \cdot \mathbf{w}_j = \mathbf{w}_i^T \mathbf{w}_j = (A\mathbf{q}_i)^T A\mathbf{q}_j = \mathbf{q}_i^T A^T A \mathbf{q}_j = \sigma_j^2 \mathbf{q}_i^T \mathbf{q}_j = \sigma_j^2 \mathbf{q}_i \cdot \mathbf{q}_j = \begin{cases} 0, & i \neq j, \\ \sigma_i^2, & i = j. \end{cases}$$

Consequently,  $\mathbf{w}_1, \dots, \mathbf{w}_r$  form an orthogonal system of vectors having respective norms

$$\|\mathbf{w}_i\| = \sqrt{\mathbf{w}_i \cdot \mathbf{w}_i} = \sigma_i.$$

We conclude that the associated unit vectors

$$\mathbf{p}_i = \frac{\mathbf{w}_i}{\sigma_i} = \frac{A\mathbf{q}_i}{\sigma_i}, \quad i = 1, \dots, r, \quad (8.55)$$

form an orthonormal set of vectors satisfying the required equations (8.53).  $Q.E.D.$

---

<sup>†</sup> Throughout this section, we exclusively use the Euclidean dot product and norm.

**Remark.** If  $A$  has distinct singular values, its singular value decomposition (8.52) is almost unique, modulo simultaneously changing the signs of one or more corresponding columns of  $Q$  and  $P$ . Matrices with repeated singular values have more freedom in their singular value decomposition, since one can choose singular vectors using different orthonormal bases of each eigenspace of the Gram matrix  $A^T A$ .

Observe that, taking the transpose of (8.52) and noting that  $\Sigma^T = \Sigma$  is diagonal, we obtain

$$A^T = Q \Sigma P, \quad (8.56)$$

which is a singular value decomposition of the transposed matrix  $A^T$ . In particular, we obtain the following result:

**Corollary 8.64.** A matrix  $A$  and its transpose  $A^T$  have the same singular values.

Note that their singular vectors are not the same; indeed, those of  $A$  are the orthonormal columns of  $Q$ , whereas those of  $A^T$  are the orthonormal columns of  $P$ , which are related by (8.53). Thus,

$$A^T \mathbf{p}_i = \sigma_i \mathbf{q}_i, \quad i = 1, \dots, r, \quad (8.57)$$

which is also a consequence of (8.54).

Furthermore, the singular value decomposition (8.52) serves to diagonalize the Gram matrix; indeed, since  $P^T P = I$ , we have

$$Q^T K Q = Q^T A^T A Q = Q^T A^T P^T P A Q = (P A Q)^T (P A Q) = \Sigma^T \Sigma = \Sigma^2, \quad (8.58)$$

because  $\Sigma$  is diagonal. If  $A$  has maximal rank  $n$ , then  $Q$  is an  $n \times n$  orthogonal matrix, and so (8.58) implies that the linear transformation defined by the Gram matrix  $K$  is diagonalized when expressed in terms of the orthonormal basis formed by the singular vectors. If  $r = \text{rank } A < n$ , then one can supplement the  $r$  singular vectors  $\mathbf{u}_1, \dots, \mathbf{u}_r$  with  $n - r$  unit vectors  $\mathbf{u}_{r+1}, \dots, \mathbf{u}_n \in \ker K = \ker A$  so as to form an orthonormal basis of  $\mathbb{R}^n$ . In terms of this basis, the Gram matrix assumes diagonal form with the  $r$  nonzero squared singular values (nonzero eigenvalues of  $K$ ) as its first  $r$  diagonal entries, while the remaining diagonal entries are all 0, in accordance with the Spectral Theorem 8.38.

**Example 8.65.** For the matrix  $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$  in Example 8.61, the orthonormal eigenbasis of  $K = A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$  is given by the unit singular vectors  $\mathbf{q}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$  and  $\mathbf{q}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ . Thus,  $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ . Next, according to (8.55),

$$\mathbf{p}_1 = \frac{A \mathbf{q}_1}{\sigma_1} = \frac{1}{\sqrt{40}} \begin{pmatrix} 4\sqrt{2} \\ 2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}, \quad \mathbf{p}_2 = \frac{A \mathbf{q}_2}{\sigma_2} = \frac{1}{\sqrt{10}} \begin{pmatrix} \sqrt{2} \\ -2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix},$$

and thus  $P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}$ . You may wish to validate the resulting singular value factorization

$$A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = P \Sigma Q^T.$$

The singular value decomposition is revealing some interesting new geometrical information concerning the action of matrices, further supplementing the discussion begun in Section 2.5 and continued in Section 4.4.

**Proposition 8.66.** Given the singular value decomposition  $A = P \Sigma Q^T$ , the columns  $\mathbf{q}_1, \dots, \mathbf{q}_r$  of  $Q$  form an orthonormal basis for  $\text{coimg } A$ , while the columns  $\mathbf{p}_1, \dots, \mathbf{p}_r$  of  $P$  form an orthonormal basis for  $\text{img } A$ .

*Proof:* The first part of the proposition was proved during the course of the proof of Theorem 8.63. Moreover, the vectors  $\mathbf{p}_i = A(\mathbf{q}_i/\sigma_i)$  for  $i = 1, \dots, r$  are mutually orthogonal, of unit length, and belong to  $\text{img } A$ , which has dimension  $r = \text{rank } A$ . They therefore form an orthonormal basis for the image. *Q.E.D.*

If  $A$  is a nonsingular  $n \times n$  matrix, then the matrices  $P, \Sigma, Q$  appearing in its singular value decomposition (8.52) are all of size  $n \times n$ . If we interpret them as linear transformations of  $\mathbb{R}^n$ , the two orthogonal matrices represent rigid rotations/reflections, while the diagonal matrix  $\Sigma$  represents a combination of simple stretches, by an amount given by the singular values, in the orthogonal coordinate directions. Thus, every invertible linear transformation on  $\mathbb{R}^n$  can be decomposed into a rotation/reflection  $Q^T$ , followed by the stretching transformation along the coordinate axes represented by  $\Sigma$ , followed by another rotation/reflection  $P$ . See also Exercise 8.5.28.

In the more general rectangular case, the matrix  $Q^T$  represents an orthogonal projection from  $\mathbb{R}^n$  to  $\text{coimg } A$ , the matrix  $\Sigma$  continues to represent a stretching transformation within this  $r$ -dimensional subspace, while  $P$  maps the result to  $\text{img } A \subset \mathbb{R}^m$ . We already noted in Section 4.4 that the linear transformation  $L: \mathbb{R}^n \rightarrow \mathbb{R}^m$  defined by matrix multiplication,  $L[\mathbf{x}] = A\mathbf{x}$ , can be interpreted as a projection from  $\mathbb{R}^n$  to  $\text{coimg } A$  followed by an invertible map from  $\text{coimg } A$  to  $\text{img } A$ . The singular value decomposition tells us that not only is the latter map invertible, it is simply a combination of stretches in the  $r$  mutually orthogonal singular directions  $\mathbf{q}_1, \dots, \mathbf{q}_r$ , whose magnitudes equal the nonzero singular values. In this way, we have at last reached a complete understanding of the subtle geometry underlying the simple operation of multiplying a vector by a matrix!

Finally, we note that practical numerical algorithms for computing singular values and the singular value decomposition can be found in Chapter 9 and in [32, 66, 90].

## The Pseudoinverse

The singular value decomposition enables us to substantially generalize the concept of a matrix inverse. The pseudoinverse was first defined by the American mathematician Eliakim Moore in the 1920's and rediscovered by the British mathematical physicist Sir Roger Penrose in the 1950's, and often has their names attached.

**Definition 8.67.** The *pseudoinverse* of a nonzero  $m \times n$  matrix with singular value decomposition  $A = P \Sigma Q^T$  is the  $n \times m$  matrix  $A^+ = Q \Sigma^{-1} P^T$ .

Note that the latter equation is the singular value decomposition of the pseudoinverse  $A^+$ , and hence its nonzero singular values are the reciprocals of the nonzero singular values of  $A$ . The only matrix without a pseudoinverse is the zero matrix  $\mathbf{0}$ . If  $A$  is a non-singular square matrix, then its pseudoinverse agrees with its ordinary inverse. Indeed, since in this case both  $P$  and  $Q$  are square orthogonal matrices, it follows that

$$A^{-1} = (P \Sigma Q^T)^{-1} = (Q^{-1})^T \Sigma^{-1} P^{-1} = Q \Sigma^{-1} P^T = A^+,$$

where we used the fact that the inverse of an orthogonal matrix is equal to its transpose. More generally, if  $A$  has linearly independent columns, or, equivalently,  $\ker A = \{\mathbf{0}\}$ , then we can bypass the singular value decomposition to compute its pseudoinverse.

**Lemma 8.68.** Let  $A$  be an  $m \times n$  matrix of rank  $n$ . Then

$$A^+ = (A^T A)^{-1} A^T. \quad (8.59)$$

*Proof:* Replacing  $A$  by its singular value decomposition (8.52), we find

$$A^T A = (P \Sigma Q^T)^T (P \Sigma Q^T) = Q \Sigma P^T P \Sigma Q^T = Q \Sigma^2 Q^T, \quad (8.60)$$

since  $\Sigma = \Sigma^T$  is a diagonal matrix, while  $P^T P = \mathbf{I}$ , since the columns of  $P$  are orthonormal. This is merely the spectral factorization (8.35) of the Gram matrix  $A^T A$  — which we in fact already knew from the original definition of the singular values and vectors. Now if  $A$  has rank  $n$ , then  $Q$  is an  $n \times n$  orthogonal matrix, and so  $Q^{-1} = Q^T$ . Therefore,

$$(A^T A)^{-1} A^T = (Q \Sigma^{-2} Q^T)^{-1} (P \Sigma Q^T)^T = (Q \Sigma^{-2} Q^T) (Q \Sigma P^T) = Q \Sigma^{-1} P^T = A^+. \quad Q.E.D.$$

If  $A$  is square and nonsingular, then, as we know, the solution to the linear system  $A \mathbf{x} = \mathbf{b}$  is given by  $\mathbf{x}^* = A^{-1} \mathbf{b}$ . For a general coefficient matrix, the vector  $\mathbf{x}^* = A^+ \mathbf{b}$  obtained by applying the pseudoinverse to the right-hand side plays a distinguished role — it is, in fact, the *least squares solution* to the system under the Euclidean norm.

**Theorem 8.69.** Consider the linear system  $A \mathbf{x} = \mathbf{b}$ . Let  $\mathbf{x}^* = A^+ \mathbf{b}$ , where  $A^+$  is the pseudoinverse of  $A$ . If  $\ker A = \{\mathbf{0}\}$ , then  $\mathbf{x}^*$  is the (Euclidean) least squares solution to the linear system. If, more generally,  $\ker A \neq \{\mathbf{0}\}$ , then  $\mathbf{x}^* = A^+ \mathbf{b} \in \text{coimg } A$  is the least squares solution that has the minimal Euclidean norm among all vectors that minimize the least squares error  $\|A \mathbf{x} - \mathbf{b}\|^2$ .

*Proof:* To show that  $\mathbf{x}^* = A^+ \mathbf{b}$  is the least squares solution to the system, we must, according to Theorem 5.11 check that it satisfies the normal equations  $A^T A \mathbf{x}^* = A^T \mathbf{b}$ . If  $\text{rank } A = n$ , so  $A^T A$  is nonsingular, this follows immediately from (8.59). More generally, combining (8.60), the definition of the pseudoinverse, and the fact that  $Q$  has orthonormal columns, so  $Q^T Q = \mathbf{I}$ , yields

$$A^T A \mathbf{x}^* = A^T A A^+ \mathbf{b} = (Q \Sigma^2 Q^T) (Q \Sigma^{-1} P^T) \mathbf{b} = Q \Sigma P^T \mathbf{b} = A^T \mathbf{b}.$$

This proves that  $\mathbf{x}^*$  solves the normal equations, and hence, by Theorem 5.11 and Exercise 5.4.11, minimizes the least squares error. Moreover,

$$\begin{aligned} \mathbf{x}^* &= A^+ \mathbf{b} = Q \Sigma^{-1} P^T \mathbf{b} \\ &= Q \mathbf{c} = c_1 \mathbf{q}_1 + \dots + c_r \mathbf{q}_r, \end{aligned} \quad \text{where} \quad \mathbf{c} = (c_1, \dots, c_r)^T = \Sigma^{-1} P^T \mathbf{b}.$$

Thus,  $\mathbf{x}^*$  is a linear combination of the singular vectors, and hence, by Proposition 8.66 and Theorem 4.50,  $\mathbf{x}^* \in \text{coimg } A$  is the solution with minimal norm; the most general least squares solution has the form  $\mathbf{x} = \mathbf{x}^* + \mathbf{z}$  for arbitrary  $\mathbf{z} \in \ker A$ . *Q.E.D.*

**Example 8.70.** Let us use the pseudoinverse to solve the linear system  $A \mathbf{x} = \mathbf{b}$ , with

$$A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & -4 & 1 \\ -1 & 3 & -1 \\ 2 & -1 & 0 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \end{pmatrix}.$$

In this case,  $\ker A \neq \{\mathbf{0}\}$ , and so we are not able use the simpler formula (8.59); thus, we begin by establishing the singular value decomposition of  $A$ . The corresponding Gram matrix

$$K = A^T A = \begin{pmatrix} 15 & -15 & 3 \\ -15 & 30 & -9 \\ 3 & -9 & 3 \end{pmatrix}$$

has eigenvalues and eigenvectors

$$\begin{aligned} \lambda_1 &= 24 + 3\sqrt{34} \simeq 41.4929, & \lambda_2 &= 24 - 3\sqrt{34} \simeq 6.5071, & \lambda_3 &= 0, \\ \mathbf{v}_1 &\simeq \begin{pmatrix} 2.1324 \\ -3.5662 \\ 1. \end{pmatrix}, & \mathbf{v}_2 &\simeq \begin{pmatrix} -2.5324 \\ -1.2338 \\ 1. \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}. \end{aligned}$$

The singular values are the square roots of the positive eigenvalues, and so

$$\sigma_1 = \sqrt{\lambda_1} \simeq 6.4415, \quad \sigma_2 = \sqrt{\lambda_2} \simeq 2.5509,$$

which are used to construct the diagonal singular value matrix  $\Sigma \simeq \begin{pmatrix} 6.4415 & 0 \\ 0 & 2.5509 \end{pmatrix}$ .

Note that  $A$  has rank 2, because it has just two singular values. The first two eigenvectors of  $K$  are the singular vectors of  $A$ , and we use the normalized (unit) singular vectors to form the columns of  $Q = (\mathbf{q}_1 \ \mathbf{q}_2) \simeq \begin{pmatrix} .4990 & -.8472 \\ -.8344 & -.4128 \\ .2340 & .3345 \end{pmatrix}$ . Next, we apply  $A$  to the singular vectors and divide by the corresponding singular value as in (8.55); the resulting vectors

$$\mathbf{p}_1 = \frac{A\mathbf{q}_1}{\sigma_1} \simeq \begin{pmatrix} -.2180 \\ .7869 \\ -.5024 \\ .2845 \end{pmatrix}, \quad \mathbf{p}_2 = \frac{A\mathbf{q}_2}{\sigma_2} \simeq \begin{pmatrix} -.7869 \\ -.2180 \\ -.2845 \\ -.5024 \end{pmatrix},$$

will form the orthonormal columns of  $P = (\mathbf{p}_1, \mathbf{p}_2)$ , and, as you can verify,

$$A = P\Sigma Q^T \simeq \begin{pmatrix} -.2180 & -.7869 \\ .7869 & -.2180 \\ -.5024 & -.2845 \\ .2845 & -.5024 \end{pmatrix} \begin{pmatrix} 6.4415 & 0 \\ 0 & 2.5509 \end{pmatrix} \begin{pmatrix} .4990 & -.8344 & .2340 \\ -.8472 & -.4128 & .3345 \end{pmatrix}$$

is the singular value decomposition of our coefficient matrix. Its pseudoinverse is immediately computed:

$$A^+ = Q\Sigma^{-1}P^T \simeq \begin{pmatrix} .2444 & .1333 & .0556 & .1889 \\ .1556 & -.0667 & .1111 & .0444 \\ -.1111 & 0 & -.0556 & -.0556 \end{pmatrix}.$$

Finally, according to Theorem 8.69, the least squares solution to the original linear system of minimal Euclidean norm is

$$\mathbf{x}^* = A^+\mathbf{b} \simeq \begin{pmatrix} .5667 \\ .1333 \\ -.1667 \end{pmatrix}.$$

## The Euclidean Matrix Norm

Singular values allow us to finally write down a formula for the natural matrix norm induced by the Euclidean norm (or 2 norm) on  $\mathbb{R}^n$ , as defined in Theorem 3.20.

**Theorem 8.71.** Let  $\|\cdot\|_2$  denote the Euclidean norm on  $\mathbb{R}^n$ . Let  $A$  be a nonzero matrix with singular values  $\sigma_1 \geq \dots \geq \sigma_r$ . Then the Euclidean matrix norm of  $A$  equals its dominant (largest) singular value:

$$\|A\|_2 = \max \{ \|A\mathbf{u}\|_2 \mid \|\mathbf{u}\|_2 = 1 \} = \sigma_1, \quad \text{while} \quad \|\mathbf{0}\|_2 = 0. \quad (8.61)$$

*Proof:* Let  $\mathbf{q}_1, \dots, \mathbf{q}_n$  be an orthonormal basis of  $\mathbb{R}^n$  consisting of the singular vectors  $\mathbf{q}_1, \dots, \mathbf{q}_r$  along with an orthonormal basis  $\mathbf{q}_{r+1}, \dots, \mathbf{q}_n$  of  $\ker A$ . Thus, by (8.57),

$$A\mathbf{q}_i = \begin{cases} \sigma_i \mathbf{p}_i, & i = 1, \dots, r, \\ 0, & i = r + 1, \dots, n, \end{cases}$$

where  $\mathbf{p}_1, \dots, \mathbf{p}_r$  form an orthonormal basis for  $\text{img } A$ . Suppose  $\mathbf{u}$  is any unit vector, so

$$\mathbf{u} = c_1 \mathbf{q}_1 + \dots + c_n \mathbf{q}_n, \quad \text{where} \quad \|\mathbf{u}\| = \sqrt{c_1^2 + \dots + c_n^2} = 1,$$

thanks to the orthonormality of the basis vectors  $\mathbf{q}_1, \dots, \mathbf{q}_n$  and the general Pythagorean formula (4.5). Then

$$A\mathbf{u} = c_1 \sigma_1 \mathbf{p}_1 + \dots + c_r \sigma_r \mathbf{p}_r, \quad \text{and hence} \quad \|A\mathbf{u}\| = \sqrt{c_1^2 \sigma_1^2 + \dots + c_r^2 \sigma_r^2},$$

since  $\mathbf{p}_1, \dots, \mathbf{p}_n$  are also orthonormal. Now, since  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ , we have

$$\begin{aligned} \|A\mathbf{u}\|_2 &= \sqrt{c_1^2 \sigma_1^2 + \dots + c_r^2 \sigma_r^2} \leq \sqrt{c_1^2 \sigma_1^2 + \dots + c_r^2 \sigma_1^2} \\ &= \sigma_1 \sqrt{c_1^2 + \dots + c_r^2} \leq \sigma_1 \sqrt{c_1^2 + \dots + c_n^2} = \sigma_1. \end{aligned}$$

Moreover, if  $c_1 = 1, c_2 = \dots = c_n = 0$ , then  $\mathbf{u} = \mathbf{q}_1$ , and hence  $\|A\mathbf{u}\|_2 = \|A\mathbf{q}_1\|_2 = \|\sigma_1 \mathbf{p}_1\|_2 = \sigma_1$ . This implies the desired formula (8.61). *Q.E.D.*

**Example 8.72.** Consider the matrix  $A = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}$ . The corresponding Gram matrix

$$A^T A \simeq \begin{pmatrix} .2225 & .0800 & .1250 \\ .0800 & .1511 & -.1111 \\ .1250 & -.1111 & .3611 \end{pmatrix},$$

has eigenvalues  $\lambda_1 \simeq .4472$ ,  $\lambda_2 \simeq .2665$ ,  $\lambda_3 \simeq .0210$ , and hence the singular values of  $A$  are their square roots:  $\sigma_1 \simeq .6687$ ,  $\sigma_2 \simeq .5163$ ,  $\sigma_3 \simeq .1448$ . The Euclidean matrix norm of  $A$  is the largest singular value, and so  $\|A\|_2 \simeq .6687$ .

## Condition Number and Rank

Not only do the singular values provide a compelling geometric interpretation of the action of a matrix on vectors, they also play a key role in modern computational algorithms. The relative magnitudes of the singular values can be used to distinguish well-behaved linear systems from ill-conditioned systems, which are more challenging to solve accurately. This information is quantified by the condition number of the matrix.

**Definition 8.73.** The *condition number* of a nonsingular  $n \times n$  matrix is the ratio between its largest and smallest singular values:  $\kappa(A) = \sigma_1/\sigma_n$ .

Since the number of singular values equals the matrix's rank, an  $n \times n$  matrix with fewer than  $n$  singular values is singular, and is said to have condition number  $\infty$ . A matrix with a very large condition number is close to singular, and designated as *ill-conditioned*; in practical terms, this occurs when the condition number is larger than the reciprocal of the machine's precision, e.g.,  $10^7$  for typical single-precision arithmetic. As the name implies, it is much harder to solve a linear system  $A\mathbf{x} = \mathbf{b}$  when its coefficient matrix is ill-conditioned and hence close to singular.

Determining the rank of a large (square or rectangular) matrix can be a numerical challenge. Small numerical inaccuracies can have an unpredictable effect. For example,

$$\text{rank 1 matrix } A = \begin{pmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 3 & 3 & -3 \end{pmatrix} \text{ is very close to } \tilde{A} = \begin{pmatrix} 1.00001 & 1. & -1. \\ 2. & 2.00001 & -2. \\ 3. & 3. & -3.00001 \end{pmatrix},$$

which has rank 3 and so is nonsingular. On the other hand, the latter matrix is very close to singular, and this is highlighted by its singular values, which are  $\sigma_1 \approx 6.48075$  while  $\sigma_2 \approx \sigma_3 \approx .000001$ . The fact that the second and third singular values are very small indicates that  $\tilde{A}$  is close to a matrix of rank 1 and should be viewed as a numerical (or experimental) perturbation of such a matrix. Thus, an effective practical method for computing the rank of a matrix is to first assign a threshold, e.g.,  $10^{-5}$ , to distinguish small singular values, and then treat any singular value lying below the threshold as if it were zero.

This idea is justified by the following theorem, which gives a mechanism for constructing the closest low rank approximations to a given matrix  $A$ , as measured in the Euclidean matrix norm.

**Theorem 8.74.** Let the  $m \times n$  matrix  $A$  have rank  $r$  and singular value decomposition  $A = P\Sigma Q^T$ . Given  $1 \leq k \leq r$ , let  $\Sigma_k$  denote the upper left  $k \times k$  diagonal submatrix of  $\Sigma$  containing the largest  $k$  singular values on its diagonal. Let  $Q_k$  denote the  $n \times k$  matrix formed from the first  $k$  columns of  $Q$ , which are the first  $k$  orthonormal singular vectors of  $A$ , and let  $P_k$  be the  $m \times k$  matrix formed from the first  $k$  columns of  $P$ . Then the  $m \times n$  matrix  $A_k = P_k \Sigma_k Q_k^T$  has rank  $k$ . Moreover,  $A_k$  is the closest rank  $k$  matrix to  $A$  in the sense that, among all  $m \times n$  matrices  $B$  of rank  $k$ , the Euclidean matrix norm  $\|A - B\|$  is minimized when  $B = A_k$ .

*Proof:* The fact that  $A_k$  has rank  $k$  is clear, since, by construction, its singular values are  $\sigma_1, \dots, \sigma_k$ . Let  $\tilde{\Sigma}_k$  denote the  $r \times r$  diagonal matrix whose first  $k$  diagonal entries are  $\sigma_1, \dots, \sigma_k$  and whose last  $r - k$  diagonal entries are all 0. Clearly  $A_k = P \tilde{\Sigma}_k Q^T$  since the additional zero entries have no effect on the product. Moreover,  $\Sigma - \tilde{\Sigma}_k$  is a diagonal matrix whose first  $k$  diagonal entries are all 0 and whose last  $r - k$  diagonal entries are  $\sigma_{k+1}, \dots, \sigma_r$ . Thus, the difference  $A - A_k = P(\Sigma - \tilde{\Sigma}_k)Q^T$  has singular values  $\sigma_{k+1}, \dots, \sigma_r$ . Since  $\sigma_{k+1}$  is the largest of these, Theorem 8.71 implies that

$$\|A - A_k\| = \sigma_{k+1}.$$

We now prove that this is the smallest possible among all  $m \times n$  matrices  $B$  of rank  $k$ . For such a matrix, according to the Fundamental Theorem 2.49,  $\dim \ker B = n - k$ . Let  $V_{k+1} \subset \mathbb{R}^n$  denote the  $(k+1)$ -dimensional subspace spanned by the first  $k+1$  singular vectors  $\mathbf{q}_1, \dots, \mathbf{q}_{k+1}$  of  $A$ . Since the dimensions of the subspaces  $V_{k+1}$  and  $\ker B$  sum up to  $k+1+n-k=n+1>n$ , their intersection is a nontrivial subspace, and hence we can

find a non-zero unit vector

$$\mathbf{u}^* = c_1 \mathbf{q}_1 + \cdots + c_{k+1} \mathbf{q}_{k+1} \in V_{k+1} \cap \ker B.$$

Thus, since  $\mathbf{q}_1, \dots, \mathbf{q}_{k+1}$  are orthonormal,

$$\|\mathbf{u}^*\|^2 = c_1^2 + \cdots + c_{k+1}^2 = 1, \quad \text{and, moreover,} \quad B\mathbf{u}^* = \mathbf{0},$$

which implies

$$(A - B)\mathbf{u}^* = A\mathbf{u}^* = c_1 A \mathbf{q}_1 + \cdots + c_{k+1} A \mathbf{q}_{k+1} = c_1 \sigma_1 \mathbf{p}_1 + \cdots + c_{k+1} \sigma_{k+1} \mathbf{p}_{k+1}.$$

Since  $\mathbf{p}_1, \dots, \mathbf{p}_{k+1}$  are also orthonormal,

$$\|(A - B)\mathbf{u}^*\|^2 = c_1^2 \sigma_1^2 + \cdots + c_{k+1}^2 \sigma_{k+1}^2 \geq (c_1^2 + \cdots + c_{k+1}^2) \sigma_{k+1}^2 = \sigma_{k+1}^2.$$

Thus, using the definition (3.39) of the Euclidean matrix norm

$$\|A - B\| = \max \{ \|(A - B)\mathbf{u}\| \mid \|\mathbf{u}\| = 1 \} \geq \|(A - B)\mathbf{u}^*\| \geq \sigma_{k+1}.$$

This proves that  $\sigma_{k+1}$  minimizes  $\|A - B\|$  among all such matrices  $B$ . *Q.E.D.*

**Remark.** One cannot do any better with a matrix of lower rank, i.e.,  $\|A - B\|$  is also minimized when  $B = A_k$  among all matrices with rank  $B \leq k$ . Justifying this statement is left to Exercise 8.7.19.

Observe that the closest rank  $k$  approximating matrix  $A_k$  is unique unless the  $(k+1)^{\text{st}}$  singular value equals the  $k^{\text{th}}$  one:  $\sigma_k = \sigma_{k+1}$ , in which case one can replace the last columns of  $P_k$  and  $Q_k$  with those coming from the  $(k+1)^{\text{st}}$  singular vectors. To compute the rank  $k$  approximating matrix  $A_k$ , we need only compute the largest  $k$  singular values  $\sigma_1, \dots, \sigma_k$  of  $A$ , which form the diagonal entries of  $\Sigma_k$ , and corresponding singular unit vectors  $\mathbf{q}_1, \dots, \mathbf{q}_k$ , which form the columns of  $Q_k$ . The columns of  $P_k$  are formed by their images  $\mathbf{p}_1 = A \mathbf{q}_1 / \sigma_1, \dots, \mathbf{p}_k = A \mathbf{q}_k / \sigma_k$ .

Consequently, when solving an ill-conditioned linear system  $A \mathbf{x} = \mathbf{b}$ , a common and effective regularization strategy is to eliminate all “insignificant” singular values below a specified cut-off, replacing  $A$  by its rank  $k$  approximation  $A_k$  specified by Theorem 8.74, where  $k$  denotes the number of significant singular values. Applying the corresponding approximating pseudoinverse  $A_k^+ = Q_k \Sigma_k^{-1} P_k^T$  to solve for  $\mathbf{x}^* = A_k^+ \mathbf{b}$  will, in favorable situations, effectively circumvent the effects of ill-conditioning.

Another common application of low rank approximations is in data compression, in which one replaces a very large data matrix, e.g., one obtained from high-resolution digital images, by a suitable low rank approximation that captures the essential features of the data set while reducing overall storage requirements and thereby accelerating subsequent analysis thereof.

## Spectral Graph Theory

Spectral graph theory, [14, 76], refers to the study of the properties of graphs that are captured by the spectrum, meaning the eigenvalues and singular values, of certain naturally associated matrices. The *graph Laplacian matrix*, which we encountered in Section 6.2, is of particular importance. Recall that it is defined as the Gram matrix,  $K = A^T A$ , constructed from the incidence matrix  $A$  of any underlying digraph, noting that the directions assigned to the edges do not affect the ultimate form of  $K$ . The eigenvalues of the graph Laplacian matrix are, by definition, the squares of the singular values of the incidence matrix.

As we know — see Exercise 2.6.12 — the dimension of the kernel of the incidence matrix, and hence also that of its graph Laplacian matrix, equals the number of connected components of the graph. In particular, a connected graph has a one-dimensional kernel spanned by the vector  $(1, 1, \dots, 1)^T$ . The magnitude of its final, meaning smallest, singular value,  $\sigma_r$ , can be interpreted as a measure of how close the graph is to being disconnected, since if it were zero (and thus technically not a singular value), the graph would have (at least) two connected components. This is borne out by numerical experiments, which demonstrate that a graph with a small final singular value  $\sigma_r$  can be disconnected by deleting a relatively small number of its edges.

**Example 8.75.** Consider the graph sketched in Figure 8.4. Using the indicated vertex labels, we can construct its graph Laplacian directly using the recipe found in Section 6.2:

$$K = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 3 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}. \quad (8.62)$$

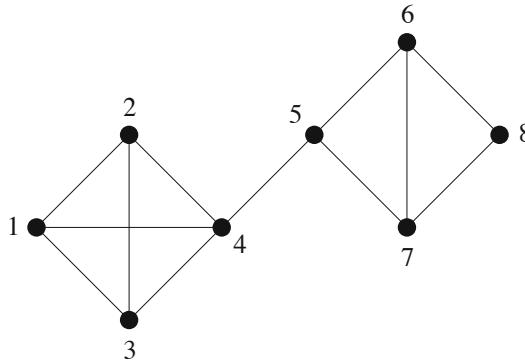
To four decimal places, the eigenvalues are 5.3234, 4., 4., 4., 4., 2.3579, .3187, 0., with corresponding singular values 2.3073, 2., 2., 2., 2., 1.5356, .5645, 0. The relatively small value of  $\sigma_7 = .5645$  indicates the graph is not especially well connected. Indeed, we can disconnect it by erasing just the one edge from vertex 4 to vertex 5. The resulting disconnected graph Laplacian is the block diagonal matrix

$$\tilde{K} = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

whose spectrum is the union of the spectra of the two constituent subgraphs, the left one having a triple eigenvalue of 4 and a zero eigenvalue, the right one having eigenvalues 4, 2, 2, 0. The singular values are their square roots. Note that these are fairly close to those of the original connected graph. Such observations are even more striking when one is dealing with much larger graphs.

Spectral graph theory plays an increasingly important role in theoretical computer science and data analysis. Applications include partitioning and coloring graphs, random graphs and random walks on graphs, routing and networks, and spectral clustering. The PageRank algorithm that underlies Google's search engine is based on representing the web pages on the internet as a gigantic digraph,<sup>†</sup> which is then viewed as a probabilistic

<sup>†</sup> According to our conventions, the internet digraph is not simple, since vertices can have two directed edges connecting them, one if the first web page links to the second and another if the second links to the first.



**Figure 8.4.** An Almost Disconnected Graph.

Markov process; see Section 9.3 for further details.

A basic example is the *complete graph*  $G_n$  on  $n$  vertices, which has one edge joining every distinct pair of vertices, and hence is the most connected simple graph; see Exercise 2.6.10. Its graph Laplacian is easily constructed, and is the  $n \times n$  matrix  $K_n = nI - E$ , where  $E$  is the  $n \times n$  matrix with every entry equal to 1. Since  $\dim \ker E = n - 1$  (why?), we see that  $K_n$  has a single nonzero eigenvalue, namely  $\lambda_1 = n$ , of multiplicity  $n - 1$  along with its zero eigenvalue. Thus, the complete graph on  $n$  vertices has  $n - 1$  identical singular values:  $\sigma_1 = \dots = \sigma_{n-1} = \sqrt{n}$ . Motivated by this observation, graphs whose nonzero singular values are close together are, in a certain sense, very highly connected, and are known as *expander graphs*. Expander graphs have many remarkable properties, which underlie their many applications, including communication networks, error-correcting codes, fault-tolerant circuits, pseudo-random number generators, Markov processes, statistical physics, as well as more theoretical disciplines such as group theory and geometry, [45].

## Exercises

- 8.7.1. Find the singular values of the following matrices: (a)  $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & -2 \\ -3 & 6 \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}$ , (e)  $\begin{pmatrix} 2 & 1 & 0 & -1 \\ 0 & -1 & 1 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$ .
- 8.7.2. Write out the singular value decomposition (8.52) of the matrices in Exercise 8.7.1.
- 8.7.3.(a) Construct the singular value decomposition of the shear matrix  $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ .  
 (b) Explain how a shear can be realized as a combination of a rotation, and a stretch, followed by a second rotation.
- 8.7.4. Find the condition number of the following matrices. Which would you characterize as ill-conditioned? (a)  $\begin{pmatrix} 2 & -1 \\ -3 & 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} -.999 & .341 \\ -1.001 & .388 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 & 2 \\ 1.001 & 1.9997 \end{pmatrix}$ , (d)  $\begin{pmatrix} -1 & 3 & 4 \\ 2 & 10 & 6 \\ 1 & 2 & -3 \end{pmatrix}$ , (e)  $\begin{pmatrix} 72 & 96 & 103 \\ 42 & 55 & 59 \\ 67 & 95 & 102 \end{pmatrix}$ , (f)  $\begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix}$ .
- 8.7.5. Find the closest rank 1 and rank 2 matrices to the matrices in Exercise 8.7.4.

- ♠ 8.7.6. Solve the following systems of equations using Gaussian Elimination with three-digit rounding arithmetic. Is your answer a reasonable approximation to the exact solution? Compare the accuracy of your answers with the condition number of the coefficient matrix, and discuss the implications of ill-conditioning.

$$(a) \begin{array}{l} 1000x + 999y = 1, \\ 554x + 555y = -1, \end{array} \quad (b) \begin{array}{l} 97x + 175y + 83z = 1, \\ 44x + 78y + 37z = 1, \\ 52x + 97y + 46z = 1. \end{array} \quad (c) \begin{array}{l} 3.001x + 2.999y + 5z = 1, \\ -x + 1.002y - 2.999z = 2 \\ 2.002x + 4y + 2z = 1.002. \end{array}$$

- ♠ 8.7.7. (a) Compute the singular values and condition numbers of the  $2 \times 2$ ,  $3 \times 3$ , and  $4 \times 4$  Hilbert matrices. (b) What is the smallest Hilbert matrix with condition number larger than  $10^6$ ?

- 8.7.8. (a) What are the singular values of a  $1 \times n$  matrix? (b) Write down its singular value decomposition. (c) Write down its pseudoinverse.

8.7.9. Answer Exercise 8.7.8 for an  $m \times 1$  matrix.

8.7.10. *True or false:* Every matrix has at least one singular value.

8.7.11. Explain why the singular values of  $A$  are the same as the nonzero eigenvalues of the positive definite square root matrix  $S = \sqrt{A^T A}$ , defined in Exercise 8.5.27.

- ◊ 8.7.12. Prove that if the square matrix  $A$  is nonsingular, then the singular values of  $A^{-1}$  are the reciprocals of the singular values of  $A$ . How are their condition numbers related?

8.7.13. *True or false:* The singular values of  $A^T$  are the same as the singular values of  $A$ .

- ◊ 8.7.14. (a) Let  $A$  be a nonsingular matrix. Prove that the product of the singular values of  $A$  equals the absolute value of its determinant:  $\sigma_1 \sigma_2 \cdots \sigma_n = |\det A|$ .  
 (b) Does their sum equal the absolute value of the trace:  $\sigma_1 + \cdots + \sigma_n = |\operatorname{tr} A|$ ?  
 (c) Show that if  $|\det A| < 10^{-k}$ , then its minimal singular value satisfies  $\sigma_n < 10^{-k/n}$ .  
 (d) *True or false:* A matrix whose determinant is very small is ill-conditioned.  
 (e) Construct an ill-conditioned matrix with  $\det A = 1$ .

8.7.15. *True or false:* If  $A$  is a symmetric matrix, then its singular values are the same as its eigenvalues.

8.7.16. *True or false:* If  $U$  is an upper triangular matrix whose diagonal entries are all positive, then its singular values are the same as its diagonal entries.

8.7.17. *True or false:* The singular values of  $A^2$  are the squares  $\sigma_i^2$  of the singular values of  $A$ .

8.7.18. *True or false:* If  $B = S^{-1}AS$  are similar matrices, then  $A$  and  $B$  have the same singular values.

- ◊ 8.7.19. Under the assumptions of Theorem 8.74, show that  $\|A - B\|_2$  is also minimized when  $B = A_k$  among all matrices with  $\operatorname{rank} B \leq k$ .

- ◊ 8.7.20. Suppose  $A$  is an  $m \times n$  matrix of rank  $r < n$ . Prove that there exist arbitrarily close matrices of maximal rank, that is, for every  $\varepsilon > 0$  there exists an  $m \times n$  matrix  $B$  with  $\operatorname{rank} B = n$  such that the Euclidean matrix norm  $\|A - B\| < \varepsilon$ .

8.7.21. *True or false:* If  $\det A > 1$ , then  $A$  is not ill-conditioned.

- 8.7.22. Let  $A = \begin{pmatrix} 6 & -4 & 1 \\ -4 & 6 & -1 \\ 1 & -1 & 11 \end{pmatrix}$ , and let  $E = \{\mathbf{y} = A\mathbf{x} \mid \|\mathbf{x}\| = 1\}$  be the image of the unit

Euclidean sphere under the linear map induced by  $A$ . (a) Explain why  $E$  is an ellipsoid and write down its equation. (b) What are its principal axes and their lengths — the semi-axes of the ellipsoid? (c) What is the volume of the solid ellipsoidal domain enclosed by  $E$ ?

◇ 8.7.23. Let  $A$  be a nonsingular  $2 \times 2$  matrix with singular value decomposition  $A = P\Sigma Q^T$  and singular values  $\sigma_1 \geq \sigma_2 > 0$ . (a) Prove that the image of the unit (Euclidean) circle under the linear transformation defined by  $A$  is an ellipse,  $E = \{A\mathbf{x} \mid \|\mathbf{x}\| = 1\}$ , whose principal axes are the columns  $\mathbf{p}_1, \mathbf{p}_2$  of  $P$ , and whose corresponding semi-axes are the singular values  $\sigma_1, \sigma_2$ . (b) Show that if  $A$  is symmetric, then the ellipse's principal axes are the eigenvectors of  $A$  and the semi-axes are the absolute values of its eigenvalues. (c) Prove that the area of  $E$  equals  $\pi |\det A|$ . (d) Find the principal axes, semi-axes, and area of the ellipses defined by (i)  $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ , (ii)  $\begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}$ , (iii)  $\begin{pmatrix} 5 & -4 \\ 0 & -3 \end{pmatrix}$ . (e) What happens if  $A$  is singular?

◇ 8.7.24. *Optimization Principles for Singular Values:* Let  $A$  be any nonzero  $m \times n$  matrix. Prove that (a)  $\sigma_1 = \max \{ \|A\mathbf{u}\| \mid \|\mathbf{u}\| = 1\}$ . (b) Is the minimum the smallest singular value? (c) Can you design optimization principles for the intermediate singular values?

◇ 8.7.25. Let  $A$  be a square matrix. Prove that its maximal eigenvalue is smaller than its maximal singular value:  $\max |\lambda_i| \leq \max \sigma_i$ . *Hint:* Use Exercise 8.7.24.

8.7.26. Compute the Euclidean matrix norm of the following matrices.

$$(a) \begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{6} \end{pmatrix}, (b) \begin{pmatrix} \frac{5}{3} & \frac{4}{3} \\ -\frac{7}{6} & -\frac{5}{6} \end{pmatrix}, (c) \begin{pmatrix} \frac{2}{7} & -\frac{2}{7} \\ -\frac{2}{7} & \frac{6}{7} \end{pmatrix}, (d) \begin{pmatrix} \frac{1}{4} & \frac{3}{2} \\ -\frac{1}{2} & \frac{5}{4} \end{pmatrix},$$

$$(e) \begin{pmatrix} \frac{2}{7} & \frac{2}{7} & -\frac{4}{7} \\ 0 & \frac{2}{7} & \frac{6}{7} \\ \frac{2}{7} & \frac{4}{7} & \frac{2}{7} \end{pmatrix}, (f) \begin{pmatrix} 0 & .1 & .8 \\ -.1 & 0 & .1 \\ -.8 & -.1 & 0 \end{pmatrix}, (g) \begin{pmatrix} 1 & -\frac{2}{3} & -\frac{2}{3} \\ 1 & -\frac{1}{3} & -1 \\ \frac{1}{3} & -\frac{2}{3} & 0 \end{pmatrix}, (h) \begin{pmatrix} \frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{pmatrix}.$$

8.7.27. Find a matrix  $A$  whose Euclidean matrix norm satisfies  $\|A^2\| \neq \|A\|^2$ .

◇ 8.7.28. Let  $K > 0$  be a positive definite matrix. Characterize the matrix norm induced by the inner product  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T K \mathbf{y}$ . *Hint:* Use Exercise 8.5.45.

8.7.29. Let  $A = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$ . Compute the matrix norm  $\|A\|$  using the following norms in  $\mathbb{R}^2$ : (a) the weighted  $\infty$  norm  $\|\mathbf{v}\| = \max\{2|v_1|, 3|v_2|\}$ ; (b) the weighted 1 norm  $\|\mathbf{v}\| = 2|v_1| + 3|v_2|$ ; (c) the weighted inner product norm  $\|\mathbf{v}\| = \sqrt{2v_1^2 + 3v_2^2}$ ; (d) the norm associated with the positive definite matrix  $K = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ .

◇ 8.7.30. Let  $A$  be an  $n \times n$  matrix with singular value vector  $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_r)$ . Prove that

(a)  $\|\boldsymbol{\sigma}\|_\infty = \|A\|_2$ ; (b)  $\|\boldsymbol{\sigma}\|_2 = \|A\|_F$ , the Frobenius norm of Exercise 3.3.51.

**Remark.** The 1 norm of the singular value vector  $\|\boldsymbol{\sigma}\|_1$  also defines a useful matrix norm, known as the *Ky Fan norm*.

◇ 8.7.31. Let  $A$  be an  $m \times n$  matrix with singular values  $\sigma_1, \dots, \sigma_r$ . Prove that

$$\sum_{i=1}^r \sigma_i^2 = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2. \quad (8.63)$$

8.7.32. Let  $A$  be a nonsingular square matrix. Prove the following formulas for its condition number:

$$(a) \kappa(A) = \frac{\max\{ \|A\mathbf{u}\| \mid \|\mathbf{u}\| = 1\}}{\min\{ \|A\mathbf{u}\| \mid \|\mathbf{u}\| = 1\}}, \quad (b) \kappa(A) = \|A\|_2 \|A^{-1}\|_2.$$

8.7.33. Find the pseudoinverse of the following matrices: (a)  $\begin{pmatrix} 1 & -1 \\ -3 & 3 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}$ ,

$$(c) \begin{pmatrix} 2 & 0 \\ 0 & -1 \\ 0 & 0 \end{pmatrix}, (d) \begin{pmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, (e) \begin{pmatrix} 1 & -1 & 1 \\ -2 & 2 & -2 \end{pmatrix}, (f) \begin{pmatrix} 1 & 3 \\ 2 & 6 \\ 3 & 9 \end{pmatrix}, (g) \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix}.$$

- 8.7.34. Use the pseudoinverse to find the least squares solution of minimal norm to the following linear systems: (a)  $x + y = 1, 3x + 3y = -2$ ; (b)  $x + y + z = 5, 2x - y + z = 2$ ; (c)  $2x + y = -1, x + y = 0$ .  $x - 3y = 2$ ,  $2x + y = -1$ .
- ♡ 8.7.35. Prove that the pseudoinverse satisfies the following identities: (a)  $(A^+)^+ = A$ , (b)  $AA^+A = A$ , (c)  $A^+AA^+ = A^+$ , (d)  $(AA^+)^T = AA^+$ , (e)  $(A^+A)^T = A^+A$ .
- 8.7.36. Suppose  $\mathbf{b} \in \text{img } A$  and  $\ker A = \{\mathbf{0}\}$ . Prove that  $\mathbf{x}^* = A^+\mathbf{b}$  is the unique solution to the linear system  $A\mathbf{x} = \mathbf{b}$ . What if  $\ker A \neq \{\mathbf{0}\}$ ?
- 8.7.37. Choose a direction for each of the edges and write down the incidence matrix  $A$  for the graph sketched in [Figure 8.4](#). Verify that its graph Laplacian (8.62) equals  $K = A^T A$ .
- 8.7.38. Determine the spectrum for the graphs in Exercise 2.6.3.
- 8.7.39. Determine the spectrum of a graph given by the edges of (i) a triangle; (ii) a square; (iii) a pentagon. Can you determine the formula for the spectrum of the graph given by an  $n$  sided polygon? *Hint:* See Exercise 8.2.49.
- 8.7.40. Determine the spectrum for the trees in Exercise 2.6.9. Can you make any conjectures about the nature of the spectrum of a graph that is a tree?

## 8.8 Principal Component Analysis

Singular values and vectors also underlie contemporary statistical data analysis. In particular, the method of *Principal Component Analysis* has assumed an increasingly essential role in a wide range of applications, including data mining, machine learning, image processing, speech recognition, semantics, face recognition, and health informatics; see [47] and the references therein. Given a large data matrix — containing many data points belonging to a high-dimensional vector space — the singular vectors associated with the larger singular values indicate the *principal components* of the data, while small singular values indicate relatively unimportant components. Projection onto the low-dimensional subspaces spanned by the dominant singular vectors can expose structure in their otherwise inscrutable large data sets. The earliest descriptions of the method of Principal Component Analysis are to be found in the first half of the twentieth century in the work of the British statistician Karl Pearson, [65], and American statistician Harold Hotelling, [46].

### Variance and Covariance

We begin with a brief description of basic statistical concepts. Suppose that  $x_1, \dots, x_m \in \mathbb{R}$  represent a collection of  $m$  measurements of a single physical quantity, e.g., the distance to a star as measured by various physical apparatuses, the speed of a car at a given instant measured by a collection of instruments, a person's IQ as measured by a series of tests, etc. Experimental error, statistical fluctuations, quantum mechanical effects, numerical approximations, and the like imply that the individual measurements will almost certainly not precisely agree. Nevertheless, one wants to know the most likely value of the measured quantity and the degree of confidence that one has in the proposed value. A variety of statistical tests have been devised to resolve these issues, and we refer the reader to, for example, [20, 43, 87].

The most basic collective quantity of such a data set is its *mean*, or *average*, denoted by

$$\bar{x} = \frac{x_1 + \dots + x_m}{m}. \quad (8.64)$$



**Figure 8.5.** Variance.

Barring some inherent statistical or experimental bias, the mean can be viewed as the most likely value, known as the *expected value*, of the quantity being measured, and thus the best bet for its actual value. (More generally, if the measurements are sampled from a known probability distribution, then one works with a suitably weighted average. To keep the formulas relatively simple, we will assume a uniform distribution throughout, and leave generalizations for the reader to pursue by consulting the statistical literature.) Once this has been computed, it will be helpful to *normalize* the measurements to have mean zero, which is done by subtracting off their mean, letting

$$a_i = x_i - \bar{x}, \quad i = 1, \dots, m, \quad \text{with} \quad \bar{a} = \frac{a_1 + \dots + a_m}{m} = 0, \quad (8.65)$$

represent the *deviations* of the measurements from their overall mean. It will also help to assemble these quantities into column vectors:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} = \mathbf{x} - \bar{x} \mathbf{e}, \quad \text{where} \quad \mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^m, \quad \bar{x} = \frac{\mathbf{e} \cdot \mathbf{x}}{m}.$$

Thus, we can write the normalized measurement vector as

$$\mathbf{a} = \left( \mathbf{I} - \frac{1}{m} \mathbf{E} \right) \mathbf{x}, \quad (8.66)$$

where  $\mathbf{I}$  is the  $m \times m$  identity matrix and  $\mathbf{E} = \mathbf{e} \mathbf{e}^T$  is the  $m \times m$  matrix all of whose entries equal 1.

The measurement *variance* tells us how widely the data points are “scattered” about their mean. As in least squares analysis, this is quantified by summing the squares of their deviations, and denoted by

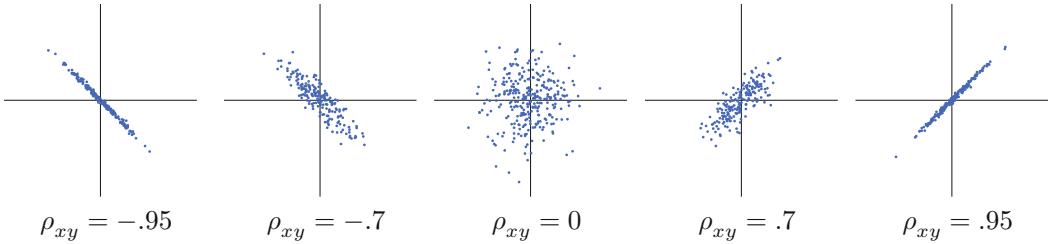
$$\sigma_x^2 = \nu \left[ (x_1 - \bar{x})^2 + \dots + (x_m - \bar{x})^2 \right] = \nu (a_1^2 + \dots + a_m^2) = \nu \|\mathbf{a}\|^2 = \sigma_a^2, \quad (8.67)$$

where we continue to use the usual Euclidean norm<sup>†</sup> throughout, and  $\nu > 0$  is a certain specified prefactor, which could also be viewed as an overall weight to the Euclidean norm. The square root of the variance is known as the *standard deviation*, and denoted by

$$\sigma = \sigma_x = \sigma_a = \sqrt{\nu} \|\mathbf{a}\|. \quad (8.68)$$

The prefactor  $\nu$  can assume different values depending upon one’s statistical objectives; common examples are (a)  $\nu = 1/m$  for the “naïve” variance; (b)  $\nu = 1/(m-1)$  (assuming  $m > 1$ , i.e., there are at least 2 measurements) for an unbiased version; (c)  $\nu = 1/(m+1)$  for the minimal mean squared estimation of variance; and (d) more exotic

<sup>†</sup> More general probability distributions rely on suitably weighted norms, which can be straightforwardly incorporated into the mathematical framework.



**Figure 8.6.** Correlation.

choices, e.g., if one desires an unbiased estimation of standard deviation instead of variance, cf. [43; p. 349]. Fortunately, apart from the resulting numerical values, the underlying analysis is independent of the prefactor.

The smaller the variance or standard deviation, the less spread out the measurements, and hence the more accurately the mean  $\bar{x}$  is expected to approximate the true value of the physical quantity. Figure 8.5 contains several *scatter plots*, in which each real-valued measurement is indicated by a dot and their mean is represented by a small vertical bar. The left plot shows data with relatively small variance, since the measurements are closely clustered about their mean, whereas on the right plot, the variance is large because the data is fairly spread out.

Now suppose we make measurements of several different physical quantities. The individual variances in themselves fail to capture many important features of the resulting data set. For example, Figure 8.6 shows the scatter plots of data sets each representing simultaneous measurements of two quantities, as specified by their horizontal and vertical coordinates. All have the same variances, both individual and cumulative, but clearly represent different interrelationships between the two quantities. In the central plot, they are completely uncorrelated, while on either side they are progressively more correlated (or anti-correlated), meaning that the value of the first measurement is a strong indicator of the value of the second.

This motivates introducing what is known as the *covariance* between a pair of measured quantities  $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$  and  $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$  as the expected value of the product of the deviations from their respective means  $\bar{x}, \bar{y}$ . We set

$$\sigma_{xy} = \nu \sum_{k=1}^m (x_k - \bar{x})(y_k - \bar{y}) = \nu \sum_{k=1}^m a_k b_k = \nu \mathbf{a} \cdot \mathbf{b}, \quad (8.69)$$

where the normalized vector  $\mathbf{a} = (a_1, a_2, \dots, a_m)^T$  has components  $a_k = x_k - \bar{x}$ , while  $\mathbf{b} = (b_1, b_2, \dots, b_m)^T$  has components  $b_k = y_k - \bar{y}$ . Note that, in view of (8.67), the covariance of a set of measurements with itself is its variance:  $\sigma_{xx} = \sigma_x^2$ . The *correlation* between the two measurement sets is then defined as

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad (8.70)$$

and is independent of the prefactor  $\nu$ . There is an overall bound on the correlation, since the Cauchy–Schwarz inequality (3.18) implies that

$$|\rho_{xy}| \leq \sigma_x \sigma_y, \quad \text{and hence} \quad -1 \leq \rho_{xy} \leq 1. \quad (8.71)$$

The closer  $\rho_{xy}$  is to  $+1$ , the more the variables are correlated; the closer to  $-1$ , the more they are anti-correlated, while  $\rho_{xy} = 0$  when the variables are uncorrelated. In Figure 8.6,

each scatter plot is labelled by its correlation. Statistically independent variables are automatically uncorrelated, but the converse is not necessarily true, since correlation measures only linear dependencies, and it is possible for nonlinearly dependent variables to nevertheless have zero correlation.

More generally, suppose we are given  $m$  measurements of  $n$  distinct quantities  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . Let  $X$  be the  $m \times n$  *data matrix* whose entry  $x_{ij}$  represents the  $i^{\text{th}}$  value or measurement of the  $j^{\text{th}}$  quantity. The column  $\mathbf{x}_j = (x_{1j}, \dots, x_{mj})^T$  of  $X$  contains the measurements of the  $j^{\text{th}}$  quantity, while the row  $\boldsymbol{\xi}_i = (x_{i1}, \dots, x_{in})$  is the  $i^{\text{th}}$  data point. For example, if the data comes from sampling a set  $S \subset \mathbb{R}^n$ , each row of the data matrix represents a different sample point  $\boldsymbol{\xi}_i \in S$ . Similarly, in image analysis, each row of the data matrix represents an individual image, whose components are, say, gray scale data for the individual pixels, or color components of pixels — in this case 3 or 4 components per pixel for the RGB or CMKY color scales — or Fourier or wavelet coefficients representing the image, etc.

The (row) vector containing the various measurement means is

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_n) = \frac{1}{m} \mathbf{e}^T X. \quad (8.72)$$

We let  $\mathbf{a}_i$ , with entries  $a_{ij} = x_{ij} - \bar{x}_j$  representing the deviations from the mean, denote the corresponding normalized (mean zero) measurement vectors, which form the columns of the *normalized data matrix*

$$A = (\mathbf{a}_1, \dots, \mathbf{a}_n) = X - \mathbf{e} \bar{\mathbf{x}} = \left( \mathbf{I} - \frac{1}{m} \mathbf{E} \right) X; \quad (8.73)$$

cf. (8.66). The fact that the columns of  $A$  all have mean zero is equivalent to the statement that  $\mathbf{e} \in \text{coker } A$ . We will call the rows  $\boldsymbol{\alpha}_i = (a_{i1}, \dots, a_{in})$  of  $A$  the *normalized data points*.

We next define the  $n \times n$  *covariance matrix*  $K$  of the data set, whose entries equal to the pairwise covariances of the individual measurements:

$$k_{ij} = \sigma_{x_i x_j} = \nu \mathbf{a}_i \cdot \mathbf{a}_j = \nu \sum_{k=1}^m a_{ki} a_{kj} = \nu \sum_{k=1}^m (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \quad i, j = 1, \dots, n. \quad (8.74)$$

The diagonal entries of  $K$  are the individual variances:  $k_{ii} = \sigma_{x_i x_i} = \sigma_{x_i}^2$ . Observe that the covariance matrix is, up to a factor, the symmetric Gram matrix for the normalized measurement vectors  $\mathbf{a}_1, \dots, \mathbf{a}_n$ , and hence

$$K = \nu A^T A. \quad (8.75)$$

Theorem 3.34 tells us that the covariance matrix is always positive semi-definite,  $K \geq 0$ , and is positive definite,  $K > 0$ , unless the columns of  $A$  are linearly dependent, or, equivalently, there is a nontrivial exact linear relationship among the normalized measurement vectors:  $c_1 \mathbf{a}_1 + \dots + c_n \mathbf{a}_n = \mathbf{0}$  with  $c_1, \dots, c_n$  not all zero — an unlikely event given the presence of measurement errors and statistical fluctuations.

**Remark.** Principal Component Analysis requires that *all*  $n$  variables be measured the *same* number,  $m$ , of times, producing a rectangular  $m \times n$  normalized data matrix  $X$ , all of whose entries are specified. Extending the analysis to missing or unavailable data is a very active area of contemporary research, which we unfortunately do not have space to examine here. We refer the interested reader to [24, 28] and the references therein.

## The Principal Components

The covariance matrix of a data set (8.75) encodes the information concerning the possible linear dependencies and interrelationships among the data. However, due to its potentially large size, it is often not easy to extract the important components and implications. Nor is visualization a good option, since the scatter plots lie in a high-dimensional space. Standard or random projections of high-dimensional data onto two- or three-dimensional subspaces give some limited insight, but the results are highly dependent on the direction of projection and tend to obscure any underlying structure. For example, projecting the data sets in Figure 8.6 onto the  $x$ - and  $y$ -axes produces more or less the same results, thereby hiding the variety of two-dimensional correlations. A more systematic approach is to locate the so-called principal components of the data, and this leads us back to the singular value decomposition of the data matrix.

The basic idea behind *Principal Component Analysis*, often abbreviated PCA, is to focus on directions in which the variance of the data is especially large. Given the  $m \times n$  normalized data matrix  $A$ , we define the first principal direction as that in which the data experiences the most variance. By “direction”, we mean a line through the origin in  $\mathbb{R}^n$ , and the variance is computed from the orthogonal projection of the data measurements onto the line. Each line is spanned by a unit vector  $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$  with  $\|\mathbf{u}\| = 1$ . (Actually, there are two unit vectors,  $\pm \mathbf{u}$ , in each line, but, as we will see, it doesn’t matter which one we choose.) The orthogonal projection of the  $i^{\text{th}}$  normalized data point  $\mathbf{a}_i = (a_{i1}, \dots, a_{in})$  onto the line spanned by  $\mathbf{u}$  is given by the projection formula (4.41), namely  $p_i = \mathbf{a}_i \mathbf{u}$ ,  $i = 1, \dots, m$ . The result is the projected measurement vector

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_m \end{pmatrix} = A \mathbf{u} = u_1 \mathbf{a}_1 + \dots + u_n \mathbf{a}_n. \quad (8.76)$$

Our goal is to maximize its variance

$$\sigma_p^2 = \nu \|\mathbf{p}\|^2 = \nu \|A \mathbf{u}\|^2 = \nu (A \mathbf{u})^T A \mathbf{u} = \nu (\mathbf{u}^T A^T A \mathbf{u}) = \nu (\mathbf{u}^T K \mathbf{u}). \quad (8.77)$$

over all possible choices of unit vector  $\mathbf{u}$ . But, ignoring the irrelevant factor  $\nu > 0$ , this is precisely the maximization problem that was solved by Theorem 8.40. We thus immediately deduce that the first principal direction is given by the dominant unit eigenvector  $\mathbf{u} = \mathbf{q}_1$  of the covariance matrix  $K = \nu A^T A$ , or, equivalently, the dominant unit singular vector of the normalized data matrix  $A$ . The maximum variance is, up to a factor, the dominant, or largest, eigenvalue, or, equivalently, the square of the dominant singular value, namely,  $\max_{\mathbf{u}} \sigma_p^2 = \nu \lambda_1 = \nu \sigma_1^2$ , while the dominant singular value is, again up to a factor, the maximal standard deviation of the projected measurements:  $\max_{\mathbf{u}} \sigma_p = \sqrt{\nu} \sigma_1$ .

The second principal direction is assumed to be orthogonal to the first, so as to avoid contaminating it with the already noted direction of maximal variance, and is to be chosen so that the variance of its projected measurements is maximized among all such orthogonal

directions. Thus, the second principal direction will maximize  $\sigma_p^2$ , as given by (8.77), over all unit vectors  $\mathbf{u}$  satisfying  $\mathbf{u} \cdot \mathbf{q}_1 = 0$ . More generally, given the first  $j-1$  principal directions  $\mathbf{q}_1, \dots, \mathbf{q}_{j-1}$ , the  $j^{\text{th}}$  principal component is in the direction  $\mathbf{u} = \mathbf{q}_j$  that maximizes the variance

$$\sigma_p^2 = \nu(\mathbf{u}^T K \mathbf{u}) \quad \text{over all vectors } \mathbf{u} \text{ satisfying } \|\mathbf{u}\| = 1, \quad \mathbf{u} \cdot \mathbf{q}_1 = \dots = \mathbf{u} \cdot \mathbf{q}_{j-1} = 0.$$

Theorem 8.42 immediately implies that  $\mathbf{q}_j$  is a unit eigenvector of  $K$  associated with its  $j^{\text{th}}$  largest eigenvalue  $\lambda_j = \sigma_j^2$ , which therefore is, up to a factor, the  $j^{\text{th}}$  principal variance. Summarizing, we have proved the Fundamental Theorem of Principal Component Analysis.

**Theorem 8.76.** The  $j^{\text{th}}$  principal direction of a normalized data matrix  $A$  is its  $j^{\text{th}}$  unit singular vector  $\mathbf{q}_j$ . The corresponding principal standard deviation  $\sqrt{\nu} \sigma_j$  is proportional to its  $j^{\text{th}}$  singular value  $\sigma_j$ .

In applications, one designates a certain number, say  $k$ , of the dominant (largest) variances  $\nu \sigma_1^2 \geq \nu \sigma_2^2 \geq \dots \geq \nu \sigma_k^2$ , as “principal” and the corresponding unit singular vectors  $\mathbf{q}_1, \dots, \mathbf{q}_k$  as the *principal directions*. The value of  $k$  depends on the user and on the application. For example, in visualization,  $k = 2$  or  $3$  in order to plot these components of the data in the plane or in space. More generally, one could specify  $k$  based on some overall size threshold, or where there is a perceived gap in the magnitudes of the variances. Another choice is such that the principal variances are those that make up some large fraction, e.g.,  $\mu = 95\%$ , of the total variance:

$$\nu(\sigma_1^2 + \dots + \nu \sigma_k^2) = \mu \left( \nu \sum_{i=1}^r \sigma_i^2 \right) = \mu \nu \sum_{i,j=1}^n a_{ij}^2 = \mu \nu \sum_{i=1}^n \sigma_{x_i}^2, \quad (8.78)$$

where the next-to-last equality follows from (8.63).

Theorem 8.74 says that, in the latter cases, the number of principal components, i.e., the number of significant singular values, will determine the approximate rank  $k$  of the covariance matrix  $K$  and hence the data matrix  $A$ , also. Thus, the normalized data (approximately) lies in a  $k$ -dimensional subspace. Further, the variance in any direction orthogonal to principal directions is relatively small and hence relatively unimportant. As a consequence, dimensional reduction by orthogonally projecting the data vectors onto the  $k$ -dimensional subspace spanned by the principal directions (singular vectors)  $\mathbf{q}_1, \dots, \mathbf{q}_k$ , serves to eliminate significant redundancies.

The coordinates of the data in the  $j^{\text{th}}$  principal direction are provided by orthogonal projection, namely the entries of the image vectors  $A \mathbf{q}_j = \sigma_j \mathbf{p}_j$ . Thus, approximating the data by its  $k$  principal components coincides with the closest rank  $k$  approximation  $A_k = P_k \Sigma_k Q_k^T$  to the data matrix  $A$  given by Theorem 8.74. Moreover, rewriting the data in terms of the *principal coordinates*, meaning those supplied by the principal directions, serves to diagonalize the principal covariance matrix  $\nu K_k = \nu A_k^T A_k$ , since, as in (8.58),

$$Q_k^T (\nu A_k^T A_k) Q_k = \nu \Sigma_k^2,$$

the result being a diagonal matrix containing the  $k$  principal variances along the diagonal. This has the important consequence that the covariance between any two principal components of the data is zero, and thus the principal components are all uncorrelated! In geometric terms, the original data tends to form an ellipsoid in the high-dimensional data space, and the principal directions are aligned with its principal semi-axis, thereby conforming to and exposing the intrinsic geometry of the data set.

Thus, to perform Principal Component Analysis on a complete data set, consisting of  $m$  measurements of  $n$  quantities, one forms the  $m \times n$  normalized data matrix  $A$  whose  $(i, j)^{\text{th}}$  entry equals the  $i^{\text{th}}$  measurement of the  $j^{\text{th}}$  variable minus the mean of the  $j^{\text{th}}$  variable. The principal directions are the first  $k$  singular vectors of  $A$ , meaning the eigenvectors of the positive (semi-)definite covariance matrix  $K = \nu A^T A$ , while the principal variances are, up to the overall factor  $\nu$ , the corresponding eigenvalues or equivalently the squares of the singular values. The principal coordinates are given by the entries of the resulting matrix  $P_k \Sigma_k$  whose columns are the image vectors  $\sigma_j \mathbf{p}_j = A \mathbf{q}_j$  specifying the measurements in the principal directions.

## Exercises

*Note:* For simplicity, take the prefactor  $\nu = 1$  unless otherwise indicated.

8.8.1. Find the mean, the variance, and the standard deviation of the following data sets:

- (a) 1.1, 1.3, 1.5, 1.55, 1.6, 1.9, 2, 2.1; (b) 2., .9, .7, 1.5, 2.6, .3, .8, 1.4; (c) -2.9, -.5, .1, -1.5, -3.6, 1.3, .4, -.7; (d) 1.1, .2, .1, .6, 1.3, -.4, -.1, .4; (e) .9, -.4, -.8, ., 1., -1.6, -1.2, -.7.

8.8.2. Find the mean, the variance, and the standard deviation of the data sets

- $\{f(x) \mid x = i/10, i = -10, \dots, 10\}$  associated with the following functions  $f(x)$ :
- (a)  $3x + 1$ , (b)  $x^2$ , (c)  $x^3 - 2x$ , (d)  $e^{-x}$ , (e)  $\tan^{-1} x$ .

8.8.3. Determine the variance and standard deviation of the *normally distributed* data points

$$\{e^{-x^2/\sigma} \mid x = i/10, i = -10, \dots, 10\} \text{ for } \sigma = 1, 2, \text{ and } 10.$$

8.8.4. Prove that  $\sigma_{xy} = \bar{xy} - \bar{x} \bar{y}$ , where  $\bar{x}$  and  $\bar{y}$  are the means of  $\{x_i\}$  and  $\{y_i\}$ , respectively, while  $\bar{xy}$  denotes the mean of the product variable  $\{x_i y_i\}$ .

◇ 8.8.5. Show that one can compute the variance of a set of measurements without reference to the mean by the following formula

$$\sigma_x^2 = \frac{\nu}{2m} \sum_{i=1}^m \sum_{j=1}^m (x_i - x_j)^2 = \frac{\nu}{m} \sum_{i < j} (x_i - x_j)^2.$$

8.8.6. Let  $A$  be an  $m \times n$  matrix that is normalized, meaning that each of its column sums is zero. Show that  $AB$ , where  $B$  is any  $n \times k$  matrix, is also normalized.

8.8.7. Given a singular value decomposition  $A = P \Sigma Q^T$ , prove that if the columns of  $A$  have zero mean, then so do the columns of  $P \Sigma$ .

♣ 8.8.8. Construct the  $5 \times 5$  covariance matrix for the 5 data sets in Exercise 8.8.1 and find its principal variances and principal directions. What do you think is the dimension of the subspace the data lies in?

♣ 8.8.9. For each of the following subsets  $S \subset \mathbb{R}^3$ , (i) Compute a fairly dense sample of data points  $z_i \in S$ ; (ii) find the principal components of your data set, using  $\mu = .95$  in the criterion in (8.78); (iii) using your principal components, estimate the dimension of the set  $S$ . Does your estimate coincide with the actual dimension? If not, explain any discrepancies.

- (a) The line segment  $S = \{(t + 1, 3t - 1, -2t)^T \mid -1 \leq t \leq 1\}$ ;
- (b) the set of points  $z$  on the three coordinate axes with Euclidean norm  $\|z\| \leq 1$ ;
- (c) the set of “probability vectors”  $S = \{(x, y, z)^T \mid 0 \leq x, y, z \leq 1, x + y + z = 1\}$ ;
- (d) the unit ball  $S = \{\|z\| \leq 1\}$  for the Euclidean norm;
- (e) the unit sphere  $S = \{\|z\| = 1\}$  for the Euclidean norm;
- (f) the unit ball  $S = \{\|z\|_\infty \leq 1\}$  for the  $\infty$  norm;
- (g) the unit sphere  $S = \{\|z\|_\infty = 1\}$  for the  $\infty$  norm.

- ♣ 8.8.10. Using the Euclidean norm, compute a fairly dense sample of points on the unit sphere  $S = \{\mathbf{x} \in \mathbb{R}^3 \mid \|\mathbf{x}\| = 1\}$ . (a) Set  $\mu = .95$  in (8.78), and then find the principal components of your data set. Do they indicate the two-dimensional nature of the sphere? If not, why not? (b) Now look at the subset of your data that is within a distance  $r > 0$  of the north pole, i.e.,  $\|\mathbf{x} - (0, 0, 1)^T\| \leq r$ , and compute its principal components. How small does  $r$  need to be to reveal the actual dimension of  $S$ ? Interpret your calculations.
- ◇ 8.8.11. Show that the first principal direction  $\mathbf{q}_1$  can be characterized as the direction of the line that minimizes the sums of the squares of its distances to the data points.
- ♡ 8.8.12. Let  $\mathbf{x}_i = (x_i, y_i)$ ,  $i = 1, \dots, m$ , be a set of data points in the plane. Suppose  $L^* \subset \mathbb{R}^2$  is the line that minimizes the sums of the squares of the distances from the data points to it, i.e.,  $\text{dist}(\mathbf{x}, L) = \sum_{i=1}^m \text{dist}(\mathbf{x}_i, L)$ , among all lines  $L \subset \mathbb{R}^2$ . (a) Prove that  $\bar{\mathbf{x}} = (\bar{x}, \bar{y}) \in L^*$ . (b) Use Exercise 8.8.11 to find  $L^*$ . (c) Apply your result to the data points in Example 5.14 and compare the resulting line  $L^*$  with the least squares line that was found there.
-



# Chapter 9

## Iteration

Iteration, meaning the repeated application of a process or function, appears in a surprisingly wide range of applications. Discrete dynamical systems, in which the time variable has been “quantized” into individual units (seconds, days, years, etc.) are modeled by iterative systems. Most numerical solution algorithms, for both linear and nonlinear systems, are based on iterative procedures. Starting with an initial guess, the successive iterates lead to closer and closer approximations to the true solution. For linear systems of equations, there are several iterative solution algorithms that can, in favorable situations, be employed as efficient alternatives to Gaussian Elimination. Iterative methods are particularly effective for solving the very large, sparse systems arising in the numerical solution of both ordinary and partial differential equations. All practical methods for computing eigenvalues and eigenvectors rely on some form of iteration. A detailed historical development of iterative methods for solving linear systems and eigenvalue problems can be found in the recent survey paper [84]. Probabilistic iterative models known as Markov chains govern basic stochastic processes and appear in genetics, population biology, scheduling, internet search, financial markets, and many more.

In this book, we will treat only iteration of linear systems. (Nonlinear iteration is of similar importance in applied mathematics and numerical analysis, and we refer the interested reader to [40, 66, 79] for details.) Linear iteration coincides with multiplication by successive powers of a matrix; convergence of the iterates depends on the magnitude of its eigenvalues. We present a variety of convergence criteria based on the spectral radius, on matrix norms, and on eigenvalue estimates provided by the Gershgorin Theorem.

We will then turn our attention to some classical iterative algorithms that can be used to accurately approximate the solutions to linear algebraic systems. The Jacobi Method is the simplest, while an evident serialization leads to the Gauss–Seidel Method. Completely general convergence criteria are hard to formulate, although convergence is assured for the important class of strictly diagonally dominant matrices that arise in many applications. A simple modification of the Gauss–Seidel Method, known as Successive Over-Relaxation (SOR), can dramatically speed up the convergence rate.

In the following Section 9.5 we discuss some practical methods for computing eigenvalues and eigenvectors of matrices. Needless to say, we completely avoid trying to solve (or even write down) the characteristic polynomial equation. The basic Power Method and its variants, which are based on linear iteration, are used to effectively approximate selected eigenvalues. To calculate the complete system of eigenvalues and eigenvectors, the remarkable  $QR$  algorithm, which relies on the Gram–Schmidt orthogonalization procedure, is the method of choice, and we include a new proof of its convergence.

The following section describes some more recent “semi-direct” iterative algorithms for finding eigenvalues and solving linear systems, that, in contrast to the classical iterative schemes, are guaranteed to eventually produce the exact solution. These are based on the idea of a Krylov subspace, spanned by the vectors generated by repeatedly multiplying an initial vector by the coefficient matrix. The Arnoldi and Lanczos algorithms are used to find a corresponding orthonormal basis for the Krylov subspaces, and thereby approximate (some of) the eigenvalues of the matrix. Two classes of solution methods are then

presented: first, the Full Orthogonalization Method (FOM) which, for a positive definite matrix, produces the powerful technique known as Conjugate Gradients (CG), of particular importance in numerical approximation of partial differential equations. The second is the recent Generalized Minimal Residual Method (GMRES), which is effectively used for solving large sparse linear systems.

The final Section 9.7 introduces the basic ideas behind wavelets, a powerful and widely used alternative to Fourier methods for signal and image processing. While slightly off topic, it provides a nice application of orthogonality and iterative techniques, and is thus a fitting end to this chapter.

## 9.1 Linear Iterative Systems

We begin with the basic definition of an iterative system of linear equations.

**Definition 9.1.** A *linear iterative system* takes the form

$$\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad \mathbf{u}^{(0)} = \mathbf{a}, \quad (9.1)$$

where the *coefficient matrix*  $T$  has size  $n \times n$ .

We will consider both real and complex systems, and so the *iterates*<sup>†</sup>  $\mathbf{u}^{(k)}$  are vectors either in  $\mathbb{R}^n$  (which assumes that the coefficient matrix  $T$  is also real) or in  $\mathbb{C}^n$ . A linear iterative system can be viewed as a discretized version of a first order system of linear ordinary differential equations, as in (8.9), in which the state of the system, as represented by the vector  $\mathbf{u}^{(k)}$ , changes at discrete time intervals, labeled by the index  $k$ .

### Scalar Systems

As usual, to study systems one begins with an in-depth analysis of the scalar version. Consider the iterative equation

$$u^{(k+1)} = \lambda u^{(k)}, \quad u^{(0)} = a, \quad (9.2)$$

where  $\lambda, a$  and the solution  $u^{(k)}$  are all real or complex scalars. The general solution to (9.2) is easily found:

$$u^{(1)} = \lambda u^{(0)} = \lambda a, \quad u^{(2)} = \lambda u^{(1)} = \lambda^2 a, \quad u^{(3)} = \lambda u^{(2)} = \lambda^3 a,$$

and, in general,

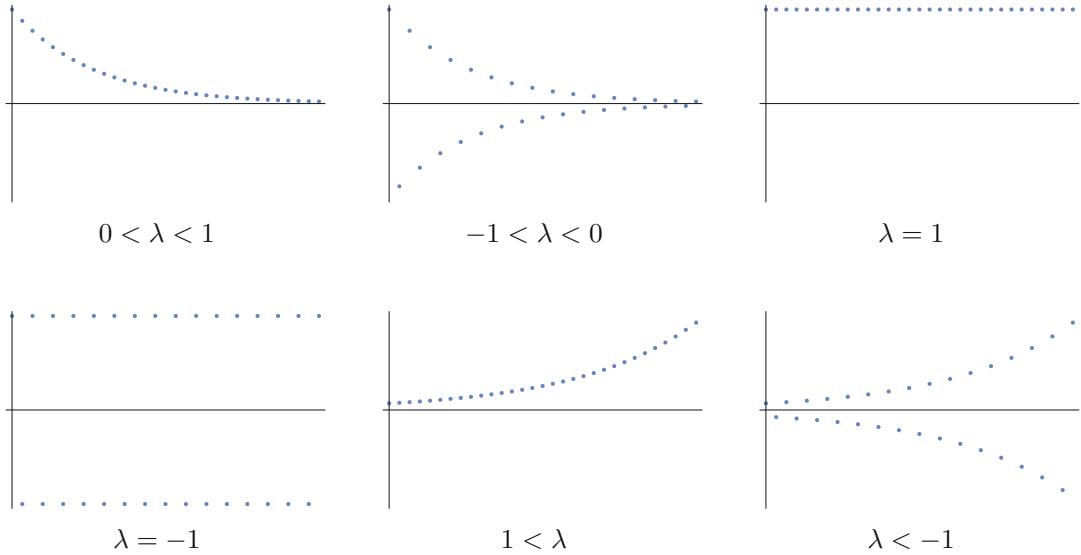
$$u^{(k)} = \lambda^k a. \quad (9.3)$$

If the initial condition is  $a = 0$ , then the solution  $u^{(k)} \equiv 0$  is constant. In other words, 0 is a *fixed point* or *equilibrium solution* for the iterative system because it does not change under iteration.

**Example 9.2.** Banks add interest to a savings account at discrete time intervals. For example, if the bank offers 5% interest compounded yearly, this means that the account balance will increase by 5% each year. Thus, assuming no deposits or withdrawals, the balance  $u^{(k)}$  after  $k$  years will satisfy the iterative equation (9.2) with  $\lambda = 1 + r$ , where

---

<sup>†</sup> **Warning.** The superscripts on  $\mathbf{u}^{(k)}$  refer to the iterate number, and should not be mistaken for derivatives.



**Figure 9.1.** One-Dimensional Real Linear Iterative Systems.

$r = .05$  is the interest rate, and the 1 indicates that all the money remains in the account. Thus, after  $k$  years, your account balance is

$$u^{(k)} = (1 + r)^k a, \quad \text{where} \quad a = u^{(0)} \quad (9.4)$$

is your initial deposit. For example, if  $u^{(0)} = a = \$1,000$ , after 1 year, your account has  $u^{(1)} = \$1,050$ , after 10 years  $u^{(10)} = \$1,628.89$ , after 50 years  $u^{(50)} = \$11,467.40$ , and after 200 years  $u^{(200)} = \$17,292,580.82$ , a gain of over 17,000%.

When the interest is compounded monthly, the rate is still quoted on a yearly basis, and so you receive  $\frac{1}{12}$  of the interest each month. If  $\hat{u}^{(k)}$  denotes the balance after  $k$  months, then, after  $n$  years, the account balance will be  $\hat{u}^{(12n)} = (1 + \frac{1}{12}r)^{12n} a$ . Thus, when the interest rate of 5% is compounded monthly, your account balance is  $\hat{u}^{(12)} = \$1,051.16$  after 1 year,  $\hat{u}^{(120)} = \$1,647.01$  after 10 years,  $\hat{u}^{(600)} = \$12,119.38$  after 50 years, and  $\hat{u}^{(2400)} = \$21,573,572.66$  dollars after 200 years. So, if you wait sufficiently long, compounding will have a dramatic effect. Similarly, daily compounding replaces 12 by 365.25, the number of days in a year. After 200 years, the balance will be  $\$22,011,396.03$ .

Let us analyze the solutions of scalar iterative equations, starting with the case when  $\lambda \in \mathbb{R}$  is a real constant. Aside from the equilibrium solution  $u^{(k)} \equiv 0$ , the iterates exhibit six qualitatively different behaviors, depending on the size of the coefficient  $\lambda$ .

- (a) If  $\lambda = 0$ , the solution immediately becomes zero, and stays there, whereby  $u^{(k)} = 0$  for all  $k \geq 1$ .
- (b) If  $0 < \lambda < 1$ , then the solution is of one sign, and tends monotonically to zero, so  $u^{(k)} \rightarrow 0$  as  $k \rightarrow \infty$ .
- (c) If  $-1 < \lambda < 0$ , then the solution tends to zero:  $u^{(k)} \rightarrow 0$  as  $k \rightarrow \infty$ . Successive iterates have alternating signs.

- (d) If  $\lambda = 1$ , the solution is constant:  $u^{(k)} = a$ , for all  $k \geq 0$ .
- (e) If  $\lambda = -1$ , the solution bounces back and forth between two values;  $u^{(k)} = (-1)^k a$ .
- (f) If  $1 < \lambda < \infty$ , then the iterates  $u^{(k)}$  become unbounded. If  $a > 0$ , they tend monotonically to  $+\infty$ ; if  $a < 0$ , they tend to  $-\infty$ .
- (g) If  $-\infty < \lambda < -1$ , then the iterates  $u^{(k)}$  also become unbounded, with alternating signs.

In [Figure 9.1](#) we exhibit representative *scatter plots* for the nontrivial cases (b – g). The horizontal axis indicates the index  $k$ , and the vertical axis the solution value  $u$ . Each dot in the scatter plot represents an iterate  $u^{(k)}$ .

In the first three cases, the fixed point  $u = 0$  is said to be *asymptotically stable*, since all solutions tend to 0 as  $k \rightarrow \infty$ . In cases (d) and (e), the zero solution is *stable*, since solutions with nearby initial data,  $|a| \ll 1$ , remain nearby. In the final two cases, the zero solution is *unstable*; every nonzero initial data  $a \neq 0$  — no matter how small — will give rise to a solution that eventually goes arbitrarily far away from equilibrium.

Let us also investigate complex scalar iterative systems. The coefficient  $\lambda$  and the initial datum  $a$  in (9.2) are allowed to be complex numbers. The solution is the same, (9.3), but now we need to know what happens when we raise a complex number  $\lambda$  to a high power. The secret is to write  $\lambda = r e^{i\theta}$  in polar form (3.93), where  $r = |\lambda|$  is its modulus and  $\theta = \text{ph } \lambda$  its angle or phase. Then  $\lambda^k = r^k e^{ik\theta}$ . Since  $|e^{ik\theta}| = 1$ , we have  $|\lambda^k| = |\lambda|^k$ , and so the solutions (9.3) have modulus  $|u^{(k)}| = |\lambda^k a| = |\lambda|^k |a|$ . As a result,  $u^{(k)}$  will remain bounded if and only if  $|\lambda| \leq 1$ , and will tend to zero as  $k \rightarrow \infty$  if and only if  $|\lambda| < 1$ .

We have thus established the basic stability criteria for scalar, linear systems.

**Theorem 9.3.** The zero solution to a (real or complex) scalar iterative system is

- (a) *asymptotically stable* if and only if  $|\lambda| < 1$ ,
- (b) *stable* if and only if  $|\lambda| \leq 1$ ,
- (c) *unstable* if and only if  $|\lambda| > 1$ .

## Exercises

- 9.1.1. Suppose  $u^{(0)} = 1$ . Find  $u^{(1)}$ ,  $u^{(10)}$ , and  $u^{(20)}$  when (a)  $u^{(k+1)} = 2u^{(k)}$ ,  
 (b)  $u^{(k+1)} = -0.9u^{(k)}$ , (c)  $u^{(k+1)} = iu^{(k)}$ , (d)  $u^{(k+1)} = (1 - 2i)u^{(k)}$ .

Is the system stable or unstable? If stable, is it asymptotically stable?

- 9.1.2. A bank offers 3.25% interest compounded yearly. Suppose you deposit \$100. (a) Set up a linear iterative equation to represent your bank balance. (b) How much money do you have after 10 years? (c) What if the interest is compounded monthly?

- 9.1.3. Show that the yearly balances of an account whose interest is compounded monthly satisfy a linear iterative system. How is the effective yearly interest rate determined from the original annual interest rate?

- 9.1.4. Show that, as the time interval of compounding goes to zero, the bank balance after  $k$  years approaches an exponential function  $e^{rk} a$ , where  $r$  is the yearly interest rate and  $a$  is the initial balance.

- 9.1.5. For which values of  $\lambda$  does the scalar iterative system (9.2) have a periodic solution, meaning that  $u^{(k+m)} = u^{(k)}$  for some  $m$ ?

9.1.6. Consider the iterative systems  $u^{(k+1)} = \lambda u^{(k)}$  and  $v^{(k+1)} = \mu v^{(k)}$ , where  $|\lambda| > |\mu|$ .

Prove that, for all nonzero initial data  $u^{(0)} = a \neq 0$ ,  $v^{(0)} = b \neq 0$ , the solution to the first is eventually larger (in modulus) than that of the second:  $|u^{(k)}| > |v^{(k)}|$ , for  $k \gg 0$ .

9.1.7. Let  $u(t)$  denote the solution to the linear ordinary differential equation  $\dot{u} = \beta u$ ,  $u(0) = a$ . Let  $h > 0$ . Show that the sample values  $u^{(k)} = u(kh)$  satisfy a linear iterative system. What is the coefficient  $\lambda$ ? Compare the stability properties of the differential equation and the corresponding iterative system.

♠ 9.1.8. Investigate the solutions of the linear iterative equation  $u^{(k+1)} = \lambda u^{(k)}$  when  $\lambda$  is a complex number with  $|\lambda| = 1$ , and look for patterns.

9.1.9. Let  $\lambda, c \in \mathbb{R}$ . Solve the *affine* or *inhomogeneous linear iterative equation*

$$u^{(k+1)} = \lambda u^{(k)} + c, \quad u^{(0)} = a. \quad (9.5)$$

Discuss the possible behaviors of the solutions. Hint: Write the solution in the form  $u^{(k)} = u^* + v^{(k)}$ , where  $u^*$  is the equilibrium solution.

9.1.10. A bank offers 5% interest compounded yearly. Suppose you deposit \$120 in the account each year. Set up an affine iterative equation (9.5) to represent your bank balance. How much money do you have after 10 years? After you retire in 50 years? After 200 years?

9.1.11. Redo Exercise 9.1.10 in the case that the interest is compounded monthly and you deposit \$10 each month.

♡ 9.1.12. Each spring, the deer in Minnesota produce offspring at a rate of roughly 1.2 times the total population, while approximately 5% of the population dies as a result of predators and natural causes. In the fall, hunters are allowed to shoot 3,600 deer. This winter the Department of Natural Resources (DNR) estimates that there are 20,000 deer. Set up an affine iterative equation (9.5) to represent the deer population each subsequent year. Solve the system and find the population in the next 5 years. How many deer in the long term will there be? Using this information, formulate a reasonable policy of how many deer hunting licenses the DNR should allow each fall, assuming one kill per license.

## Powers of Matrices

The solution to the general linear iterative system

$$\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad \mathbf{u}^{(0)} = \mathbf{a}, \quad (9.6)$$

is also, at least at first glance, immediate. Clearly,

$$\mathbf{u}^{(1)} = T \mathbf{u}^{(0)} = T \mathbf{a}, \quad \mathbf{u}^{(2)} = T \mathbf{u}^{(1)} = T^2 \mathbf{a}, \quad \mathbf{u}^{(3)} = T \mathbf{u}^{(2)} = T^3 \mathbf{a},$$

and, in general,

$$\mathbf{u}^{(k)} = T^k \mathbf{a}. \quad (9.7)$$

Thus, the iterates are simply determined by multiplying the initial vector  $\mathbf{a}$  by the successive powers of the coefficient matrix  $T$ . And so, in contrast to differential equations, proving the existence and uniqueness of solutions to an iterative system is completely trivial.

However, unlike real or complex scalars, the general formulas for and qualitative behavior of the powers of a square matrix are not nearly so immediately apparent. (Before continuing, the reader is urged to experiment with simple  $2 \times 2$  matrices, trying to detect patterns.) To make progress, recall how, in Section 8.1, we endeavored to solve linear systems of differential equations by suitably adapting the known exponential solution from the scalar version. In the iterative case, the scalar solution formula (9.3) is written in terms of powers, not exponentials. This motivates us to try the power ansatz

$$\mathbf{u}^{(k)} = \lambda^k \mathbf{v}, \quad (9.8)$$

in which  $\lambda$  is a scalar and  $\mathbf{v}$  a vector, as a possible solution to the system. We find

$$\mathbf{u}^{(k+1)} = \lambda^{k+1} \mathbf{v}, \quad \text{while} \quad T \mathbf{u}^{(k)} = T(\lambda^k \mathbf{v}) = \lambda^k T \mathbf{v}.$$

These two expressions will be equal if and only if

$$T \mathbf{v} = \lambda \mathbf{v}.$$

This is precisely the defining eigenvalue equation (8.12), and thus, (9.8) is a nontrivial solution to (9.6) if and only if  $\lambda$  is an *eigenvalue* of the coefficient matrix  $T$  and  $\mathbf{v} \neq \mathbf{0}$  an associated *eigenvector*.

Thus, for each eigenvector and eigenvalue of the coefficient matrix, we can construct a solution to the iterative system. We can then appeal to linear superposition, as in Theorem 7.30, to combine the basic eigensolutions to form more general solutions. In particular, if the coefficient matrix is complete, this method will produce the general solution.

**Theorem 9.4.** If the coefficient matrix  $T$  is complete, then the general solution to the linear iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$  is given by

$$\mathbf{u}^{(k)} = c_1 \lambda_1^k \mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2 + \cdots + c_n \lambda_n^k \mathbf{v}_n, \quad (9.9)$$

where  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are the linearly independent eigenvectors and  $\lambda_1, \dots, \lambda_n$  the corresponding eigenvalues of  $T$ . The coefficients  $c_1, \dots, c_n$  are arbitrary scalars and are uniquely prescribed by the initial conditions  $\mathbf{u}^{(0)} = \mathbf{a}$ .

*Proof:* Since we already know, by linear superposition, that (9.9) is a solution to the system for arbitrary  $c_1, \dots, c_n$ , it suffices to show that we can match any prescribed initial conditions. To this end, we need to solve the linear system

$$\mathbf{u}^{(0)} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \mathbf{a}. \quad (9.10)$$

Completeness of  $T$  implies that its eigenvectors form a basis of  $\mathbb{C}^n$ , and hence (9.10) always admits a solution. In matrix form, we can rewrite (9.10) as

$$S \mathbf{c} = \mathbf{a}, \quad \text{so that} \quad \mathbf{c} = S^{-1} \mathbf{a}, \quad \text{where} \quad S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$$

is the (nonsingular) matrix whose columns are the eigenvectors. *Q.E.D.*

Solutions in the incomplete cases are more complicated to write down, and rely on the Jordan bases of Section 8.6; see Exercise 9.1.40.

**Example 9.5.** Consider the iterative system

$$x^{(k+1)} = \frac{3}{5} x^{(k)} + \frac{1}{5} y^{(k)}, \quad y^{(k+1)} = \frac{1}{5} x^{(k)} + \frac{3}{5} y^{(k)}, \quad (9.11)$$

with initial conditions

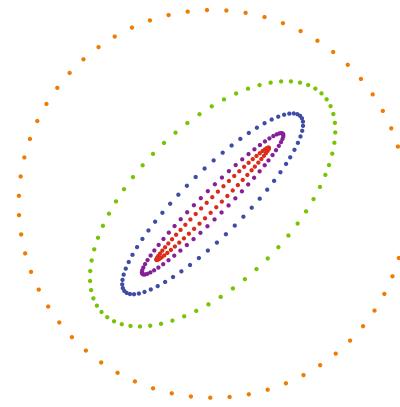
$$x^{(0)} = a, \quad y^{(0)} = b. \quad (9.12)$$

The system can be rewritten in our matrix form (9.6), with

$$T = \begin{pmatrix} .6 & .2 \\ .2 & .6 \end{pmatrix}, \quad \mathbf{u}^{(k)} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Solving the characteristic equation

$$\det(T - \lambda I) = \lambda^2 - 1.2\lambda - .32 = 0$$



**Figure 9.2.** Stable Iterative System.

produces the eigenvalues  $\lambda_1 = .8$ ,  $\lambda_2 = .4$ . We then solve the associated linear systems  $(T - \lambda_j \mathbf{I})\mathbf{v}_j = \mathbf{0}$  for the corresponding eigenvectors:

$$\lambda_1 = .8, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = .4, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Therefore, the basic eigensolutions are

$$\mathbf{u}_1^{(k)} = (.8)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{u}_2^{(k)} = (.4)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Theorem 9.4 tells us that the general solution is given as a linear combination,

$$\mathbf{u}^{(k)} = c_1 \mathbf{u}_1^{(k)} + c_2 \mathbf{u}_2^{(k)} = c_1 (.8)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 (.4)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 (.8)^k - c_2 (.4)^k \\ c_1 (.8)^k + c_2 (.4)^k \end{pmatrix},$$

where  $c_1, c_2$  are determined by the initial conditions:

$$\mathbf{u}^{(0)} = \begin{pmatrix} c_1 - c_2 \\ c_1 + c_2 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \text{and hence} \quad c_1 = \frac{a+b}{2}, \quad c_2 = \frac{b-a}{2}.$$

Therefore, the explicit formula for the solution to the initial value problem (9.11–12) is

$$x^{(k)} = (.8)^k \frac{a+b}{2} + (.4)^k \frac{a-b}{2}, \quad y^{(k)} = (.8)^k \frac{a+b}{2} + (.4)^k \frac{b-a}{2}.$$

In particular, as  $k \rightarrow \infty$ , the iterates  $\mathbf{u}^{(k)} \rightarrow \mathbf{0}$  converge to zero at a rate governed by the dominant eigenvalue  $\lambda_1 = .8$ . Figure 9.2 illustrates the cumulative effect of the iteration; the initial data is colored orange, and successive iterates are colored green, blue, purple, red. The initial conditions consist of a large number of points on the unit circle  $x^2 + y^2 = 1$ , which are successively mapped to points on progressively smaller and flatter ellipses, that shrink down towards the origin.

**Example 9.6.** The *Fibonacci numbers* are defined by the second order<sup>†</sup> scalar iterative equation

$$u^{(k+2)} = u^{(k+1)} + u^{(k)}, \tag{9.13}$$

<sup>†</sup> In general, an iterative system  $\mathbf{u}^{(k+j)} = T_1 \mathbf{u}^{(k+j-1)} + \cdots + T_j \mathbf{u}^{(k)}$  in which the new iterate depends upon the preceding  $j$  values is said to have *order  $j$* .

with initial conditions

$$u^{(0)} = a, \quad u^{(1)} = b. \quad (9.14)$$

In short, to obtain the next Fibonacci number, add the previous two. The classical *Fibonacci integers* start with  $a = 0, b = 1$ ; the next few are

$$u^{(0)} = 0, \quad u^{(1)} = 1, \quad u^{(2)} = 1, \quad u^{(3)} = 2, \quad u^{(4)} = 3, \quad u^{(5)} = 5, \quad u^{(6)} = 8, \quad u^{(7)} = 13, \quad \dots$$

The Fibonacci integers occur in a surprising variety of natural objects, including leaves, flowers, and fruit, [83]. They were originally introduced by the eleventh-/twelfth-century Italian mathematician Leonardo Pisano Fibonacci as a crude model of the growth of a population of rabbits. In Fibonacci's model, the  $k^{\text{th}}$  Fibonacci number  $u^{(k)}$  measures the total number of pairs of rabbits at year  $k$ . We start the process with a single juvenile pair<sup>‡</sup> at year 0. Once a year, each pair of rabbits produces a new pair of offspring, but it takes a full year for a rabbit pair to mature enough to produce offspring of their own.

Every higher order iterative equation can be replaced by an equivalent first order iterative system. In this particular case, we define the vector

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix} \in \mathbb{R}^2,$$

and note that (9.13) is equivalent to the matrix system

$$\begin{pmatrix} u^{(k+1)} \\ u^{(k+2)} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix}, \quad \text{or} \quad \mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad \text{where} \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

To find the explicit formula for the Fibonacci numbers, we must determine the eigenvalues and eigenvectors of the coefficient matrix  $T$ . A straightforward computation produces

$$\begin{aligned} \lambda_1 &= \frac{1 + \sqrt{5}}{2} = 1.618034\dots, & \lambda_2 &= \frac{1 - \sqrt{5}}{2} = -.618034\dots, \\ \mathbf{v}_1 &= \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}. \end{aligned}$$

Therefore, according to (9.9), the general solution to the Fibonacci system is

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix} = c_1 \left( \frac{1 + \sqrt{5}}{2} \right)^k \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \left( \frac{1 - \sqrt{5}}{2} \right)^k \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}. \quad (9.15)$$

The initial data

$$\mathbf{u}^{(0)} = c_1 \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

uniquely specifies the coefficients

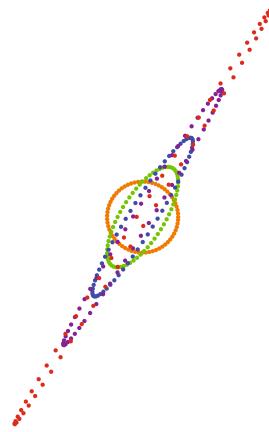
$$c_1 = \frac{2a + (1 + \sqrt{5})b}{2\sqrt{5}}, \quad c_2 = -\frac{2a + (1 - \sqrt{5})b}{2\sqrt{5}}.$$

The first entry of the solution vector (9.15) produces the explicit formula

$$u^{(k)} = \frac{(-1 + \sqrt{5})a + 2b}{2\sqrt{5}} \left( \frac{1 + \sqrt{5}}{2} \right)^k + \frac{(1 + \sqrt{5})a - 2b}{2\sqrt{5}} \left( \frac{1 - \sqrt{5}}{2} \right)^k \quad (9.16)$$

---

<sup>‡</sup> Fibonacci ignores some pertinent details like the sex of the offspring.



**Figure 9.3.** Fibonacci Iteration.

for the  $k^{\text{th}}$  Fibonacci number. For the particular initial conditions  $a = 0$ ,  $b = 1$ , (9.16) reduces to the classical *Binet formula*

$$u^{(k)} = \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^k - \left( \frac{1 - \sqrt{5}}{2} \right)^k \right]. \quad (9.17)$$

It is a remarkable fact that, for every value of  $k$ , all the  $\sqrt{5}$ 's cancel out, and the Binet formula (9.17) does indeed produce the Fibonacci integers listed above. Another useful observation is that, since

$$0 < |\lambda_2| = \frac{\sqrt{5} - 1}{2} < 1 < \lambda_1 = \frac{1 + \sqrt{5}}{2},$$

the terms involving  $\lambda_1^k$  go to  $\infty$  (and so the zero solution to this iterative system is unstable) while the terms involving  $\lambda_2^k$  go to zero. Therefore, even for  $k$  moderately large, the first term in (9.16) is an excellent approximation to the  $k^{\text{th}}$  Fibonacci number — and one that gets more and more accurate as  $k$  increases. A plot of the first 4 iterates, starting with the initial data consisting of equally spaced points on the unit circle, appears in Figure 9.3. As in the previous example, the circle is mapped to a sequence of progressively more eccentric ellipses; however, their major semi-axes become more and more stretched out, and almost all points end up going off to  $\infty$  in the direction of the dominant eigenvector  $\mathbf{v}_2$ .

The dominant eigenvalue  $\lambda_1 = \frac{1}{2}(1 + \sqrt{5}) = 1.6180339\dots$  is known as the *golden ratio* and plays an important role in spiral growth in nature, as well as in art, architecture, and design, [83]. It describes the overall growth rate of the Fibonacci integers, and, in fact, every sequence of Fibonacci numbers with initial conditions  $b \neq \frac{1}{2}(1 - \sqrt{5})a$ .

**Example 9.7.** Let  $T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix}$  be the coefficient matrix for a three-dimensional iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$ . Its eigenvalues and corresponding eigenvectors are

$$\begin{aligned} \lambda_1 &= -2, & \lambda_2 &= -1 + i, & \lambda_3 &= -1 - i, \\ \mathbf{v}_1 &= \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 2 - i \\ -1 \\ 1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 2 + i \\ -1 \\ 1 \end{pmatrix}. \end{aligned}$$

Therefore, according to (9.9), the general complex solution is

$$\mathbf{u}^{(k)} = b_1 (-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + b_2 (-1 + i)^k \begin{pmatrix} 2 - i \\ -1 \\ 1 \end{pmatrix} + b_3 (-1 - i)^k \begin{pmatrix} 2 + i \\ -1 \\ 1 \end{pmatrix},$$

where  $b_1, b_2, b_3$  are arbitrary complex scalars.

If we are interested only in real solutions, we can break up any complex solution into its real and imaginary parts, each of which constitutes a real solution. (This is a manifestation of the general Reality Principle of Theorem 7.48, but is not hard to prove directly.) We begin by writing  $\lambda_2 = -1 + i = \sqrt{2} e^{3\pi i/4}$  in polar form, and hence

$$(-1 + i)^k = 2^{k/2} e^{3k\pi i/4} = 2^{k/2} \left( \cos \frac{3}{4}k\pi + i \sin \frac{3}{4}k\pi \right).$$

Therefore, the complex solution

$$(-1 + i)^k \begin{pmatrix} 2 - i \\ -1 \\ 1 \end{pmatrix} = 2^{k/2} \begin{pmatrix} 2 \cos \frac{3}{4}k\pi + \sin \frac{3}{4}k\pi \\ -\cos \frac{3}{4}k\pi \\ \cos \frac{3}{4}k\pi \end{pmatrix} + i 2^{k/2} \begin{pmatrix} 2 \sin \frac{3}{4}k\pi - \cos \frac{3}{4}k\pi \\ -\sin \frac{3}{4}k\pi \\ \sin \frac{3}{4}k\pi \end{pmatrix}$$

is a combination of two independent real solutions. The complex conjugate eigenvalue  $\lambda_3 = -1 - i$  leads, as before, to the complex conjugate solution — and the same two real solutions. The general real solution  $\mathbf{u}^{(k)}$  to the system can be written as a linear combination of the three independent real solutions:

$$c_1 (-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + c_2 2^{k/2} \begin{pmatrix} 2 \cos \frac{3}{4}k\pi + \sin \frac{3}{4}k\pi \\ -\cos \frac{3}{4}k\pi \\ \cos \frac{3}{4}k\pi \end{pmatrix} + c_3 2^{k/2} \begin{pmatrix} 2 \sin \frac{3}{4}k\pi - \cos \frac{3}{4}k\pi \\ -\sin \frac{3}{4}k\pi \\ \sin \frac{3}{4}k\pi \end{pmatrix}, \quad (9.18)$$

where  $c_1, c_2, c_3$  are arbitrary real scalars, uniquely prescribed by the initial conditions.

## Diagonalization and Iteration

An alternative, equally efficient approach to solving iterative systems is based on diagonalization of the coefficient matrix, cf. (8.30). Specifically, assuming the coefficient matrix  $T$  is complete, we can factor it as a product

$$T = S \Lambda S^{-1}, \quad (9.19)$$

in which  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$  is the diagonal matrix containing the eigenvalues of  $T$ , while the columns of  $S = (\mathbf{v}_1 \ \dots \ \mathbf{v}_n)$  are the corresponding eigenvectors. Consequently, the powers of  $T$  are given by

$$\begin{aligned} T^2 &= (S \Lambda S^{-1})(S \Lambda S^{-1}) = S \Lambda^2 S^{-1}, \\ T^3 &= (S \Lambda S^{-1})(S \Lambda S^{-1})(S \Lambda S^{-1}) = S \Lambda^3 S^{-1}, \end{aligned}$$

and, in general,

$$T^k = S \Lambda^k S^{-1}. \quad (9.20)$$

Moreover, since  $\Lambda$  is a diagonal matrix, its powers are trivial to compute:

$$\Lambda^k = \text{diag}(\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k). \quad (9.21)$$

Thus, by combining (9.20–21), we obtain an explicit formula for the powers of a complete matrix  $T$ . Furthermore, the solution to the associated linear iterative system

$$\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad \mathbf{u}^{(0)} = \mathbf{a}, \quad \text{is given by} \quad \mathbf{u}^{(k)} = T^k \mathbf{a} = S \Lambda^k S^{-1} \mathbf{a}. \quad (9.22)$$

You should convince yourself that this gives precisely the same solution as before. Computationally, there is not a significant difference between the two solution methods, and the choice is left to the discretion of the user.

**Example 9.8.** Suppose  $T = \begin{pmatrix} 7 & 6 \\ -9 & -8 \end{pmatrix}$ . Its eigenvalues and eigenvectors are readily computed:

$$\lambda_1 = -2, \quad \mathbf{v}_1 = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \quad \lambda_2 = 1, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We assemble these into the diagonal eigenvalue matrix  $\Lambda$  and the eigenvector matrix  $S$ , given by

$$\Lambda = \begin{pmatrix} -2 & 0 \\ 0 & 1 \end{pmatrix}, \quad S = \begin{pmatrix} -2 & -1 \\ 3 & 1 \end{pmatrix},$$

whence

$$\begin{pmatrix} 7 & 6 \\ -9 & -8 \end{pmatrix} = T = S \Lambda S^{-1} = \begin{pmatrix} -2 & -1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} -2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -3 & -2 \end{pmatrix},$$

as you can readily check. Therefore, according to (9.20),

$$\begin{aligned} T^k &= S \Lambda^k S^{-1} \\ &= \begin{pmatrix} -2 & -1 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} (-2)^k & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -3 & -2 \end{pmatrix} = \begin{pmatrix} 3 - 2(-2)^k & 2 - 2(-2)^k \\ -3 + 3(-2)^k & -2 + 3(-2)^k \end{pmatrix}. \end{aligned}$$

You may wish to check this formula directly for the first few values of  $k = 1, 2, \dots$ . As a result, the solution to the particular iterative system

$$\mathbf{u}^{(k+1)} = \begin{pmatrix} 7 & 6 \\ -9 & -8 \end{pmatrix} \mathbf{u}^{(k)}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{is} \quad \mathbf{u}^{(k)} = T^k \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 - 4(-2)^k \\ -5 + 6(-2)^k \end{pmatrix}.$$

In this case, the eigenvalue  $\lambda_1 = -2$  causes an instability, with solutions having arbitrarily large norm as  $k \rightarrow \infty$ .

## Exercises

9.1.13. Find the explicit formula for the solution to the following linear iterative systems:

- $u^{(k+1)} = u^{(k)} - 2v^{(k)}, \quad v^{(k+1)} = -2u^{(k)} + v^{(k)}, \quad u^{(0)} = 1, \quad v^{(0)} = 0.$
- $u^{(k+1)} = u^{(k)} - \frac{2}{3}v^{(k)}, \quad v^{(k+1)} = \frac{1}{2}u^{(k)} - \frac{1}{6}v^{(k)}, \quad u^{(0)} = -2, \quad v^{(0)} = 3.$
- $u^{(k+1)} = u^{(k)} - v^{(k)}, \quad v^{(k+1)} = -u^{(k)} + 5v^{(k)}, \quad u^{(0)} = 1, \quad v^{(0)} = 0.$
- $u^{(k+1)} = \frac{1}{2}u^{(k)} + v^{(k)}, \quad v^{(k+1)} = v^{(k)} - 2w^{(k)}, \quad w^{(k+1)} = \frac{1}{3}w^{(k)},$   
 $u^{(0)} = 1, \quad v^{(0)} = -1, \quad w^{(0)} = 1.$
- $u^{(k+1)} = -u^{(k)} + 2v^{(k)} - w^{(k)}, \quad v^{(k+1)} = -6u^{(k)} + 7v^{(k)} - 4w^{(k)},$   
 $w^{(k+1)} = -6u^{(k)} + 6v^{(k)} - 4w^{(k)}, \quad u^{(0)} = 0, \quad v^{(0)} = 1, \quad w^{(0)} = 3.$

9.1.14. Find the explicit formula for the general solution to the linear iterative systems with the following coefficient matrices:

- $\begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$ ,
- $\begin{pmatrix} -2 & 7 \\ -1 & 3 \end{pmatrix}$ ,
- $\begin{pmatrix} -3 & 2 & -2 \\ -6 & 4 & -3 \\ 12 & -6 & -5 \end{pmatrix}$ ,
- $\begin{pmatrix} -\frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ 0 & -\frac{1}{2} & \frac{1}{3} \\ 1 & -1 & \frac{2}{3} \end{pmatrix}$ .

9.1.15. Prove that all the Fibonacci integers  $u^{(k)}$ ,  $k \geq 0$ , can be found by just computing the first term in the Binet formula (9.17) and then rounding off to the nearest integer.

9.1.16. The  $k^{\text{th}}$  *Lucas number* is defined as  $L^{(k)} = \left(\frac{1+\sqrt{5}}{2}\right)^k + \left(\frac{1-\sqrt{5}}{2}\right)^k$ .

- (a) Explain why the Lucas numbers satisfy the Fibonacci iterative equation  $L^{(k+2)} = L^{(k+1)} + L^{(k)}$ . (b) Write down the first 7 Lucas numbers.
- (c) Prove that every Lucas number is a positive integer.

9.1.17. What happens to the Fibonacci integers  $u^{(k)}$  if we go “backward in time”, i.e., for  $k < 0$ ? How is  $u^{(-k)}$  related to  $u^{(k)}$ ?

9.1.18. Use formula (9.20) to compute the  $k^{\text{th}}$  power of the following matrices:

$$(a) \begin{pmatrix} 5 & 2 \\ 2 & 2 \end{pmatrix}, \quad (b) \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}, \quad (e) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 2 \end{pmatrix}.$$

9.1.19. Use your answer from Exercise 9.1.18 to solve the following iterative systems:

- (a)  $u^{(k+1)} = 5u^{(k)} + 2v^{(k)}$ ,  $v^{(k+1)} = 2u^{(k)} + 2v^{(k)}$ ,  $u^{(0)} = -1$ ,  $v^{(0)} = 0$ ,
- (b)  $u^{(k+1)} = 4u^{(k)} + v^{(k)}$ ,  $v^{(k+1)} = -2u^{(k)} + v^{(k)}$ ,  $u^{(0)} = 1$ ,  $v^{(0)} = -3$ ,
- (c)  $u^{(k+1)} = u^{(k)} + v^{(k)}$ ,  $v^{(k+1)} = -u^{(k)} + v^{(k)}$ ,  $u^{(0)} = 0$ ,  $v^{(0)} = 2$ ,
- (d)  $u^{(k+1)} = u^{(k)} + v^{(k)} + 2w^{(k)}$ ,  $v^{(k+1)} = u^{(k)} + 2v^{(k)} + w^{(k)}$ ,  
 $w^{(k+1)} = 2u^{(k)} + v^{(k)} + w^{(k)}$ ,  $u^{(0)} = 1$ ,  $v^{(0)} = 0$ ,  $w^{(0)} = 1$ ,
- (e)  $u^{(k+1)} = v^{(k)}$ ,  $v^{(k+1)} = w^{(k)}$ ,  $w^{(k+1)} = -u^{(k)} + 2w^{(k)}$ ,  $u^{(0)} = 1$ ,  $v^{(0)} = 0$ ,  $w^{(0)} = 0$ .

9.1.20. (a) Given initial data  $\mathbf{u}^{(0)} = (1, 1, 1)^T$ , explain why the resulting solution  $\mathbf{u}^{(k)}$  to the system in Example 9.7 has all integer entries. (b) Find the coefficients  $c_1, c_2, c_3$  in the explicit solution formula (9.18). (c) Check the first few iterates to convince yourself that the solution formula does, in spite of appearances, always give an integer value.

9.1.21. (a) Show how to convert the higher order linear iterative equation

$$u^{(k+j)} = c_1 u^{(k+j-1)} + c_2 u^{(k+j-2)} + \cdots + c_j u^{(k)}$$

into a first order system  $\mathbf{u}^{(k)} = T\mathbf{u}^{(k)}$ . Hint: See Example 9.6.

(b) Write down initial conditions that guarantee a unique solution  $u^{(k)}$  for all  $k \geq 0$ .

9.1.22. Apply the method of Exercise 9.1.21 to solve the following iterative equations:

- (a)  $u^{(k+2)} = -u^{(k+1)} + 2u^{(k)}$ ,  $u^{(0)} = 1$ ,  $u^{(1)} = 2$ .
- (b)  $12u^{(k+2)} = u^{(k+1)} + u^{(k)}$ ,  $u^{(0)} = -1$ ,  $u^{(1)} = 2$ .
- (c)  $u^{(k+2)} = 4u^{(k+1)} + u^{(k)}$ ,  $u^{(0)} = 1$ ,  $u^{(1)} = -1$ .
- (d)  $u^{(k+2)} = 2u^{(k+1)} - 2u^{(k)}$ ,  $u^{(0)} = 1$ ,  $u^{(1)} = 3$ .
- (e)  $u^{(k+3)} = 2u^{(k+2)} + u^{(k+1)} - 2u^{(k)}$ ,  $u^{(0)} = 0$ ,  $u^{(1)} = 2$ ,  $u^{(2)} = 3$ .
- (f)  $u^{(k+3)} = u^{(k+2)} + 2u^{(k+1)} - 2u^{(k)}$ ,  $u^{(0)} = 0$ ,  $u^{(1)} = 1$ ,  $u^{(2)} = 1$ .

9.1.23. Suppose you have  $n$  dollars and can buy coffee for \$1, milk for \$2, and orange juice for \$2. Let  $C^{(n)}$  count the number of different ways of spending all your money. (a) Explain why  $C^{(n)} = C^{(n-1)} + 2C^{(n-2)}$ ,  $C^{(0)} = C^{(1)} = 1$ . (b) Find an explicit formula for  $C^{(n)}$ .

9.1.24. Find the general solution to the iterative system  $u_i^{(k+1)} = u_{i-1}^{(k)} + u_{i+1}^{(k)}$ ,  $i = 1, \dots, n$ , where we set  $u_0^{(k)} = u_{n+1}^{(k)} = 0$  for all  $k$ . Hint: Use Exercise 8.2.47.

- 9.1.25. Starting with  $u^{(0)} = 0$ ,  $u^{(1)} = 0$ ,  $u^{(2)} = 1$ , define the sequence of *tribonacci numbers*  $u^{(k)}$  by adding the previous three to get the next one. For instance,  
 $u^{(3)} = u^{(0)} + u^{(1)} + u^{(2)} = 1$ . (a) Write out the next four tribonacci numbers. (b) Find a third order iterative equation for the  $k^{\text{th}}$  tribonacci number. (c) Explain why the tribonacci numbers are all integers. (d) Find an explicit formula for the solution, using a computer to approximate the eigenvalues. (e) Do they grow faster than the usual Fibonacci numbers? What is their overall rate of growth?

- ♣ 9.1.26. Suppose that Fibonacci's rabbits live for only eight years, [44]. (a) Write out an iterative equation to describe the rabbit population. (b) Write down the first few terms. (c) Convert your equation into a first order iterative system, using the method of Exercise 9.1.21. (d) At what rate does the rabbit population grow?
- ♣ 9.1.27. A well-known method of generating a sequence of "pseudo-random" integers  $u^{(0)}, u^{(1)}, u^{(2)}, \dots$  satisfying  $0 \leq u^{(i)} < n$  is based on the *modular Fibonacci equation*  $u^{(k+2)} = u^{(k+1)} + u^{(k)} \bmod n$ , with suitably chosen initial values  $0 \leq u^{(0)}, u^{(1)} < n$ .  
 (a) Generate the sequence of pseudo-random numbers that result from the choices  $n = 10$ ,  $u^{(0)} = 3$ ,  $u^{(1)} = 7$ . Keep iterating until the sequence starts repeating.  
 (b) Experiment with other sequences of pseudo-random numbers generated by the method.
- 9.1.28. Prove that the curves  $E_k = \{ T^k \mathbf{x} \mid \|\mathbf{x}\| = 1 \}$ ,  $k = 0, 1, 2, \dots$ , sketched in Figure 9.2 form a family of ellipses with the same principal axes. What are the individual semi-axes?  
*Hint:* Use Exercise 8.7.23.
- ♣ 9.1.29. Plot the ellipses  $E_k = \{ T^k \mathbf{x} \mid \|\mathbf{x}\| = 1 \}$  for  $k = 1, 2, 3, 4$  for the following matrices  $T$ . Then determine their principal axes, semi-axes, and areas. *Hint:* Use Exercise 8.7.23.
- $$(a) \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & -1.2 \\ .4 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix}.$$
- 9.1.30. Let  $T$  be a positive definite  $2 \times 2$  matrix. Let  $E_n = \{ T^n \mathbf{x} \mid \|\mathbf{x}\| = 1 \}$ ,  $n = 0, 1, 2, \dots$ , be the image of the unit circle under the  $n^{\text{th}}$  power of  $T$ . (a) Prove that  $E_n$  is an ellipse.  
*True or false:* (b) The ellipses  $E_n$  all have the same principal axes. (c) The semi-axes are given by  $r_n = r_1^n$ ,  $s_n = s_1^n$ . (d) The areas are given by  $A_n = \pi \alpha^n$  where  $\alpha = A_1/\pi$ .
- 9.1.31. Answer Exercise 9.1.30 when  $T$  is an arbitrary nonsingular  $2 \times 2$  matrix.  
*Hint:* Use Exercise 8.7.23.
- 9.1.32. Given the general solution (9.9) of the iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$ , write down the solution to  $\mathbf{v}^{(k+1)} = \alpha T \mathbf{v}^{(k)} + \beta \mathbf{v}^{(k)}$ , where  $\alpha, \beta \in \mathbb{R}$ .
- ◊ 9.1.33. Prove directly that if the coefficient matrix of a linear iterative system is real, both the real and imaginary parts of a complex solution are real solutions.
- ◊ 9.1.34. Explain why the solution  $\mathbf{u}^{(k)}$ ,  $k \geq 0$ , to the initial value problem (9.6) exists and is uniquely defined. Does this hold if we allow negative  $k < 0$ ?
- 9.1.35. Prove that if  $T$  is a symmetric matrix, then the coefficients in (9.9) are given by the formula  $c_j = \mathbf{a}^T \mathbf{v}_j / \mathbf{v}_j^T \mathbf{v}_j$ .
- 9.1.36. Explain why the  $j^{\text{th}}$  column  $\mathbf{c}_j^{(k)}$  of the matrix power  $T^k$  satisfies the linear iterative system  $\mathbf{c}_j^{(k+1)} = T \mathbf{c}_j^{(k)}$  with initial data  $\mathbf{c}_j^{(0)} = \mathbf{e}_j$ , the  $j^{\text{th}}$  standard basis vector.
- 9.1.37. Let  $z^{(k+1)} = \lambda z^{(k)}$  be a complex scalar iterative equation with  $\lambda = \mu + i\nu$ . Show that its real and imaginary parts  $x^{(k)} = \operatorname{Re} z^{(k)}$ ,  $y^{(k)} = \operatorname{Im} z^{(k)}$ , satisfy a two-dimensional real linear iterative system. Use the eigenvalue method to solve the real  $2 \times 2$  system, and verify that your solution coincides with the solution to the original complex equation.
- ◊ 9.1.38. Suppose  $V \subset \mathbb{R}^n$  is an invariant subspace for the  $n \times n$  matrix  $T$  governing the linear iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$ . Prove that if  $\mathbf{u}^{(0)} \in V$ , then so is the solution:  $\mathbf{u}^{(k)} \in V$ .
- 9.1.39. Suppose  $\mathbf{u}^{(k)}$  and  $\tilde{\mathbf{u}}^{(k)}$  are two solutions to the same iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$ .  
 (a) Suppose  $\mathbf{u}^{(k_0)} = \tilde{\mathbf{u}}^{(k_0)}$  for some  $k_0 \geq 0$ . Can you conclude that these are the same solution:  $\mathbf{u}^{(k)} = \tilde{\mathbf{u}}^{(k)}$  for all  $k$ ? (b) What can you say if  $\mathbf{u}^{(k_0)} = \tilde{\mathbf{u}}^{(k_1)}$  for  $k_0 \neq k_1$ ?

- ◇ 9.1.40. Let  $T$  be an incomplete matrix, and suppose  $\mathbf{w}_1, \dots, \mathbf{w}_j$  is a Jordan chain associated with an incomplete eigenvalue  $\lambda$ . (a) Prove that, for  $i = 1, \dots, j$ ,

$$T^k \mathbf{w}_i = \lambda^k \mathbf{w}_i + k \lambda^{k-1} \mathbf{w}_{i-1} + \binom{k}{2} \lambda^{k-2} \mathbf{w}_{i-2} + \dots . \quad (9.23)$$

- (b) Explain how to use a Jordan basis of  $T$  to construct the general solution to the linear iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$ .

- 9.1.41. Use the method Exercise 9.1.40 to find the general real solution to the following linear iterative systems:

- (a)  $u^{(k+1)} = 2u^{(k)} + 3v^{(k)}$ ,  $v^{(k+1)} = 2v^{(k)}$ ,  
 (b)  $u^{(k+1)} = u^{(k)} + v^{(k)}$ ,  $v^{(k+1)} = -4u^{(k)} + 5v^{(k)}$ ,  
 (c)  $u^{(k+1)} = -u^{(k)} + v^{(k)} + w^{(k)}$ ,  $v^{(k+1)} = -v^{(k)} + w^{(k)}$ ,  $w^{(k+1)} = -w^{(k)}$ ,  
 (d)  $u^{(k+1)} = 3u^{(k)} - v^{(k)}$ ,  $v^{(k+1)} = -u^{(k)} + 3v^{(k)} + w^{(k)}$ ,  $w^{(k+1)} = -v^{(k)} + 3w^{(k)}$ ,  
 (e)  $u^{(k+1)} = u^{(k)} - v^{(k)} - w^{(k)}$ ,  $v^{(k+1)} = 2u^{(k)} + 2v^{(k)} + 2w^{(k)}$ ,  $w^{(k+1)} = -u^{(k)} + v^{(k)} + w^{(k)}$ ,  
 (f)  $u^{(k+1)} = v^{(k)} + z^{(k)}$ ,  $v^{(k+1)} = -u^{(k)} + w^{(k)}$ ,  $w^{(k+1)} = z^{(k)}$ ,  $z^{(k+1)} = -w^{(k)}$ .

- 9.1.42. Find a formula for the  $k^{\text{th}}$  power of a Jordan block matrix. Hint: Use Exercise 9.1.40.

- ◇ 9.1.43. An *affine iterative system* has the form  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)} + \mathbf{b}$ ,  $\mathbf{u}^{(0)} = \mathbf{c}$ .

- (a) Under what conditions does the system have an equilibrium solution  $\mathbf{u}^{(k)} \equiv \mathbf{u}^*$ ?  
 (b) In such cases, find a formula for the general solution. Hint: Look at  $\mathbf{v}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^*$ .  
 (c) Solve the following affine iterative systems:

$$(i) \quad \mathbf{u}^{(k+1)} = \begin{pmatrix} 6 & 3 \\ -3 & -4 \end{pmatrix} \mathbf{u}^{(k)} + \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} 4 \\ -3 \end{pmatrix},$$

$$(ii) \quad \mathbf{u}^{(k+1)} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix} \mathbf{u}^{(k)} + \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$(iii) \quad \mathbf{u}^{(k+1)} = \begin{pmatrix} -3 & 2 & -2 \\ -6 & 4 & -3 \\ 12 & -6 & -5 \end{pmatrix} \mathbf{u}^{(k)} + \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix},$$

$$(iv) \quad \mathbf{u}^{(k+1)} = \begin{pmatrix} -\frac{5}{6} & \frac{1}{3} & -\frac{1}{6} \\ 0 & -\frac{1}{2} & \frac{1}{3} \\ 1 & -1 & \frac{2}{3} \end{pmatrix} \mathbf{u}^{(k)} + \begin{pmatrix} \frac{1}{6} \\ -\frac{1}{3} \\ -\frac{1}{2} \end{pmatrix}, \quad \mathbf{u}^{(0)} = \begin{pmatrix} \frac{1}{6} \\ -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}.$$

- (d) Discuss what happens in cases in which there is no fixed point, assuming that  $T$  is complete.

## 9.2 Stability

With the solution formula (9.9) in hand, we are now in a position to understand the qualitative behavior of solutions to (complete) linear iterative systems. The most important case for applications is when all the iterates converge to  $\mathbf{0}$ .

**Definition 9.9.** The equilibrium solution  $\mathbf{u}^* = \mathbf{0}$  to a linear iterative system (9.1) is called *globally asymptotically stable* if all solutions  $\mathbf{u}^{(k)} \rightarrow \mathbf{0}$  as  $k \rightarrow \infty$ .

Asymptotic stability relies on the following property of the coefficient matrix.

**Definition 9.10.** A matrix  $T$  is called *convergent* if its powers converge to the zero matrix,  $T^k \rightarrow \mathbf{0}$ , meaning that the individual entries of  $T^k$  all go to 0 as  $k \rightarrow \infty$ .

The equivalence of the convergence condition and stability of the iterative system follows

immediately from the solution formula (9.7).

**Theorem 9.11.** The linear iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$  has globally asymptotically stable zero solution if and only if  $T$  is a convergent matrix.

*Proof:* If  $T^k \rightarrow \mathbf{0}$ , and  $\mathbf{u}^{(k)} = T^k \mathbf{a}$  is any solution, then clearly  $\mathbf{u}^{(k)} \rightarrow \mathbf{0}$  as  $k \rightarrow \infty$ , proving stability. Conversely, the solution  $\mathbf{u}_j^{(k)} = T^k \mathbf{e}_j$  is the same as the  $j^{\text{th}}$  column of  $T^k$ . If the origin is asymptotically stable, then  $\mathbf{u}_j^{(k)} \rightarrow \mathbf{0}$ . Thus, the individual columns of  $T^k$  all tend to  $\mathbf{0}$ , proving that  $T^k \rightarrow \mathbf{0}$ . *Q.E.D.*

To facilitate the analysis of convergence, we shall adopt a norm  $\|\cdot\|$  on our underlying vector space,  $\mathbb{R}^n$  or  $\mathbb{C}^n$ . The reader may be inclined to choose the Euclidean (or Hermitian) norm, but, in practice, the  $\infty$  norm

$$\|\mathbf{u}\|_\infty = \max\{ |u_1|, \dots, |u_n| \} \quad (9.24)$$

prescribed by the vector's maximal entry (in modulus) is often easier to work with. Convergence of the iterates is equivalent to convergence of their norms:

$$\mathbf{u}^{(k)} \rightarrow \mathbf{0} \quad \text{if and only if} \quad \|\mathbf{u}^{(k)}\| \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty.$$

The fundamental stability criterion for linear iterative systems relies on the size of the eigenvalues of the coefficient matrix.

**Theorem 9.12.** The matrix  $T$  is convergent, and hence the zero solution of the associated linear iterative system (9.1) is globally asymptotically stable, if and only if all its (complex) eigenvalues have modulus strictly less than one:  $|\lambda_j| < 1$ .

*Proof:* Let us prove this result assuming that the coefficient matrix  $T$  is complete. (The proof in the incomplete case relies on the Jordan canonical form, and is outlined in Exercise 9.2.18.) If  $\lambda_j$  is an eigenvalue such that  $|\lambda_j| < 1$ , then the corresponding basis solution  $\mathbf{u}_j^{(k)} = \lambda_j^k \mathbf{v}_j$  tends to zero as  $k \rightarrow \infty$ ; indeed,

$$\|\mathbf{u}_j^{(k)}\| = \|\lambda_j^k \mathbf{v}_j\| = |\lambda_j|^k \|\mathbf{v}_j\| \rightarrow 0, \quad \text{since} \quad |\lambda_j| < 1.$$

Therefore, if all eigenvalues are less than 1 in modulus, all terms in the solution formula (9.9) tend to zero, which proves asymptotic stability:  $\mathbf{u}^{(k)} \rightarrow \mathbf{0}$ . Conversely, if any eigenvalue satisfies  $|\lambda_j| \geq 1$ , then the solution  $\mathbf{u}^{(k)} = \lambda_j^k \mathbf{v}_j$  does not tend to  $\mathbf{0}$  as  $k \rightarrow \infty$ , and hence  $\mathbf{0}$  is not asymptotically stable. *Q.E.D.*

## Spectral Radius

Consequently, the necessary and sufficient condition for asymptotic stability of a linear iterative system is that all the eigenvalues of the coefficient matrix lie strictly inside the unit circle in the complex plane:  $|\lambda_j| < 1$ . This criterion can be recast using the following important definition.

**Definition 9.13.** The *spectral radius* of a matrix  $T$  is defined as the maximal modulus of all of its real and complex eigenvalues:  $\rho(T) = \max \{ |\lambda_1|, \dots, |\lambda_k| \}$ .

**Theorem 9.14.** The matrix  $T$  is convergent if and only if its spectral radius is strictly less than one:  $\rho(T) < 1$ .

If  $T$  is complete, then we can apply the triangle inequality to (9.9) to estimate

$$\begin{aligned}\|\mathbf{u}^{(k)}\| &= \|c_1 \lambda_1^k \mathbf{v}_1 + \cdots + c_n \lambda_n^k \mathbf{v}_n\| \\ &\leq |\lambda_1|^k \|c_1 \mathbf{v}_1\| + \cdots + |\lambda_n|^k \|c_n \mathbf{v}_n\| \\ &\leq \rho(T)^k (|c_1| \|\mathbf{v}_1\| + \cdots + |c_n| \|\mathbf{v}_n\|) = C \rho(T)^k,\end{aligned}\quad (9.25)$$

for some constant  $C > 0$  that depends only upon the initial conditions. In particular, if  $\rho(T) < 1$ , then

$$\|\mathbf{u}^{(k)}\| \leq C \rho(T)^k \rightarrow 0 \quad \text{as } k \rightarrow \infty, \quad (9.26)$$

in accordance with Theorem 9.14. Thus, the spectral radius  $\rho(T)$  prescribes the rate of convergence of the solutions to equilibrium; the smaller the spectral radius, the faster the solutions go to  $\mathbf{0}$ .

If  $T$  has only one largest (simple) eigenvalue, so  $|\lambda_1| > |\lambda_j|$  for all  $j > 1$ , then the first term in the solution formula (9.9) will eventually dominate all the others:  $\|\lambda_1^k \mathbf{v}_1\| \gg \|\lambda_j^k \mathbf{v}_j\|$  for  $j > 1$  and  $k \gg 0$ . Therefore, provided that  $c_1 \neq 0$ , the solution (9.9) has the asymptotic formula

$$\mathbf{u}^{(k)} \approx c_1 \lambda_1^k \mathbf{v}_1, \quad (9.27)$$

and so most solutions end up parallel to  $\mathbf{v}_1$ . In particular, if  $|\lambda_1| = \rho(T) < 1$ , such a solution approaches  $\mathbf{0}$  along the direction of the dominant eigenvector  $\mathbf{v}_1$  at a rate governed by the modulus of the dominant eigenvalue. The exceptional solutions, with  $c_1 = 0$ , tend to  $\mathbf{0}$  at a faster rate, along one of the other eigendirections. In practical computations, one rarely observes the exceptional solutions. Indeed, even if the initial condition does not involve the dominant eigenvector, numerical errors during the iteration will almost inevitably introduce a small component in the direction of  $\mathbf{v}_1$ , which will, if you wait long enough, eventually dominate the solution.

The inequality (9.25) applies only to complete matrices. In the general case, one can prove, cf. Exercise 9.2.18, that the solution satisfies the slightly weaker inequality

$$\|\mathbf{u}^{(k)}\| \leq C \sigma^k \quad \text{for all } k \geq 0, \quad \text{where } \sigma > \rho(T) \quad (9.28)$$

is any number larger than the spectral radius, while  $C > 0$  is a positive constant (whose value may depend on how close  $\sigma$  is to  $\rho$ ).

**Example 9.15.** According to Example 9.7, the matrix

$$T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix} \quad \text{has eigenvalues} \quad \begin{aligned}\lambda_1 &= -2, \\ \lambda_2 &= -1 + i, \\ \lambda_3 &= -1 - i.\end{aligned}$$

Since  $|\lambda_1| = 2 > |\lambda_2| = |\lambda_3| = \sqrt{2}$ , the spectral radius is  $\rho(T) = |\lambda_1| = 2$ . We conclude that  $T$  is not a convergent matrix. As the reader can check, either directly, or from the solution formula (9.18), the vectors  $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)}$  obtained by repeatedly multiplying any nonzero initial vector  $\mathbf{u}^{(0)}$  by  $T$  rapidly go off to  $\infty$ , in successively opposite directions, at a rate roughly equal to  $\rho(T)^k = 2^k$ .

On the other hand, the matrix

$$\tilde{T} = -\frac{1}{3} T = \begin{pmatrix} 1 & -\frac{1}{3} & -2 \\ -\frac{1}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix} \quad \text{with eigenvalues} \quad \begin{aligned}\lambda_1 &= \frac{2}{3}, \\ \lambda_2 &= \frac{1}{3} - \frac{1}{3}i, \\ \lambda_3 &= \frac{1}{3} + \frac{1}{3}i,\end{aligned}$$

has spectral radius  $\rho(\tilde{T}) = \frac{2}{3}$ , and hence is a convergent matrix. According to (9.27), if we write the initial data  $\mathbf{u}^{(0)} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3$  as a linear combination of the eigenvectors, then, provided  $c_1 \neq 0$ , the iterates have the asymptotic form  $\mathbf{u}^{(k)} \approx c_1 \left(\frac{2}{3}\right)^k \mathbf{v}_1$ , where  $\mathbf{v}_1 = (4, -2, 1)^T$  is the eigenvector corresponding to the dominant eigenvalue  $\lambda_1 = \frac{2}{3}$ . Thus, for most initial vectors, the iterates end up decreasing in length by a factor of almost exactly  $\frac{2}{3}$ , eventually becoming parallel to the dominant eigenvector  $\mathbf{v}_1$ . This is borne out by a sample computation: starting with  $\mathbf{u}^{(0)} = (1, 1, 1)^T$ , we obtain, for instance,

$$\mathbf{u}^{(15)} = \begin{pmatrix} -.018216 \\ .009135 \\ -.004567 \end{pmatrix}, \quad \mathbf{u}^{(16)} = \begin{pmatrix} -.012126 \\ .006072 \\ -.003027 \end{pmatrix}, \quad \mathbf{u}^{(17)} = \begin{pmatrix} -.008096 \\ .004048 \\ -.002018 \end{pmatrix},$$

which form progressively more accurate scalar multiples of the dominant eigenvector  $\mathbf{v}_1 = (4, -2, 1)^T$ ; moreover, the ratios between their successive entries,  $\mathbf{u}_i^{(k+1)} / \mathbf{u}_i^{(k)}$ , are approaching the dominant eigenvalue  $\lambda_1 = \frac{2}{3}$ .

## Exercises

9.2.1. Determine the spectral radius of the following matrices:

$$(a) \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad (b) \begin{pmatrix} \frac{1}{3} & -\frac{1}{4} \\ \frac{1}{2} & -\frac{1}{3} \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & 1 & 2 \end{pmatrix}, \quad (d) \begin{pmatrix} -1 & 5 & -9 \\ 4 & 0 & -1 \\ 4 & -4 & 3 \end{pmatrix}.$$

9.2.2. Determine whether or not the following matrices are convergent:

$$(a) \begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix}, \quad (b) \begin{pmatrix} .6 & .3 \\ .3 & .7 \end{pmatrix}, \quad (c) \frac{1}{5} \begin{pmatrix} 5 & -3 & -2 \\ 1 & -2 & 1 \\ 1 & -5 & 4 \end{pmatrix}, \quad (d) \begin{pmatrix} .8 & .3 & .2 \\ .1 & .2 & .6 \\ .1 & .5 & .2 \end{pmatrix}.$$

9.2.3. Which of the listed coefficient matrices defines a linear iterative system with asymptotically stable zero solution?

$$(a) \begin{pmatrix} -3 & 0 \\ -4 & -1 \end{pmatrix}, \quad (b) \begin{pmatrix} \frac{1}{2} & \frac{3}{4} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}, \quad (c) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad (d) \begin{pmatrix} -1 & 3 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & -1 \end{pmatrix},$$

$$(e) \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & -\frac{1}{4} \\ \frac{1}{2} & \frac{3}{4} & -\frac{1}{2} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix}, \quad (f) \begin{pmatrix} 3 & 0 & -1 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{pmatrix}, \quad (g) \begin{pmatrix} 1 & 0 & -3 & -2 \\ -\frac{1}{2} & \frac{1}{2} & 2 & \frac{3}{2} \\ -\frac{1}{6} & 0 & \frac{3}{2} & \frac{2}{3} \\ \frac{2}{3} & 0 & -3 & -\frac{5}{3} \end{pmatrix}.$$

9.2.4. (a) Determine the eigenvalues and spectral radius of the matrix  $T = \begin{pmatrix} 3 & 2 & -2 \\ -2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$ .

(b) Use part (a) to find the eigenvalues and spectral radius of  $\hat{T} = \begin{pmatrix} \frac{3}{5} & \frac{2}{5} & -\frac{2}{5} \\ -\frac{2}{5} & \frac{1}{5} & 0 \\ 0 & \frac{2}{5} & \frac{1}{5} \end{pmatrix}$ .

(c) Write down an asymptotic formula for the solutions to  $\mathbf{u}^{(k+1)} = \hat{T} \mathbf{u}^{(k)}$ .

9.2.5. (a) Show that the spectral radius of  $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$  is  $\rho(T) = 1$ .

(b) Show that most iterates  $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)}$  become unbounded as  $k \rightarrow \infty$ .

(c) Discuss why the inequality  $\|\mathbf{u}^{(k)}\| \leq C \rho(T)^k$  does not hold when the coefficient matrix is incomplete. (d) Can you prove that (9.28) holds in this example?

- 9.2.6. Given a linear iterative system with non-convergent matrix, which solutions, if any, will converge to  $\mathbf{0}$ ?
- ◊ 9.2.7. Suppose  $T$  is a complete matrix. (a) Prove that every solution to the corresponding linear iterative system is bounded if and only if  $\rho(T) \leq 1$ . (b) Can you generalize this result to incomplete matrices? *Hint:* Look at Exercise 9.1.40.
- ◊ 9.2.8. Discuss the asymptotic behavior of solutions to an iterative system that has two eigenvalues of largest modulus, e.g.,  $\lambda_1 = -\lambda_2$ , or  $\lambda_1 = \overline{\lambda_2}$  are complex conjugate eigenvalues. How would you detect this? How can you determine the eigenvalues and eigenvectors?
- 9.2.9. Suppose  $T$  has spectral radius  $\rho(T)$ . Can you predict the spectral radius of  $aT + bI$ , where  $a, b$  are scalars? If not, what additional information do you need?
- 9.2.10. Prove that if  $A$  is any square matrix, then there exists  $c \neq 0$  such that the scalar multiple  $cA$  is a convergent matrix. Find a formula for the largest possible such  $c$ .
- ◊ 9.2.11. Let  $M_n$  be the  $n \times n$  tridiagonal matrix with all 1's on the sub- and super-diagonals, and zeros on the main diagonal. (a) What is the spectral radius of  $M_n$ ? *Hint:* Use Exercise 8.2.47. (b) Is  $M_n$  convergent? (c) Find the general solution to the iterative system  $\mathbf{u}^{(k+1)} = M_n \mathbf{u}^{(k)}$ .
- ◊ 9.2.12. Let  $\alpha, \beta$  be scalars. Let  $T_{\alpha, \beta}$  be the  $n \times n$  tridiagonal matrix that has all  $\alpha$ 's on the sub- and super-diagonals, and  $\beta$ 's on the main diagonal. (a) Solve the iterative system  $\mathbf{u}^{(k+1)} = T_{\alpha, \beta} \mathbf{u}^{(k)}$ . (b) For which values of  $\alpha, \beta$  is the system asymptotically stable? *Hint:* Combine Exercises 9.2.11 and 9.1.32.
- 9.2.13. (a) Prove that if  $|\det T| > 1$ , then the iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$  is unstable. (b) If  $|\det T| < 1$ , is the system asymptotically stable? Prove or give a counterexample.
- 9.2.14. *True or false:* (a)  $\rho(cA) = c\rho(A)$ , (b)  $\rho(S^{-1}AS) = \rho(A)$ , (c)  $\rho(A^2) = \rho(A)^2$ , (d)  $\rho(A^{-1}) = 1/\rho(A)$ , (e)  $\rho(A+B) = \rho(A) + \rho(B)$ , (f)  $\rho(AB) = \rho(A)\rho(B)$ .
- 9.2.15. *True or false:* (a) If  $T$  is convergent, then  $T^2$  is convergent. (b) If  $A$  is convergent, then  $T = A^T A$  is convergent.
- 9.2.16. Suppose  $T^k \rightarrow P$  as  $k \rightarrow \infty$ . (a) Prove that  $P$  is idempotent:  $P^2 = P$ . (b) Can you characterize all such matrices  $P$ ? (c) What are the conditions on the matrix  $A$  for this to happen?
- 9.2.17. Prove that a matrix  $T$  with all integer entries is convergent if and only if it is nilpotent, i.e.,  $T^k = \mathbf{0}$  for some  $k \geq 0$ . Give a nonzero example of such a matrix.
- ◊ 9.2.18. Prove the inequality (9.28) when  $T$  is incomplete. Use it to complete the proof of Theorem 9.14 in the incomplete case. *Hint:* Use Exercises 9.1.40, 9.2.22.
- ◊ 9.2.19. Suppose that  $M$  is a nonsingular matrix. (a) Prove that the *implicit iterative system*  $M \mathbf{u}^{(n+1)} = \mathbf{u}^{(n)}$  has globally asymptotically stable zero solution if and only if all the eigenvalues of  $M$  are strictly greater than one in magnitude:  $|\mu_i| > 1$ . (b) Let  $K$  be another matrix. Prove that more general implicit iterative system of the form  $M \mathbf{u}^{(n+1)} = K \mathbf{u}^{(n)}$  has globally asymptotically stable zero solution if and only if all the generalized eigenvalues of the matrix pair  $K, M$ , as in Exercise 8.5.8, are strictly less than 1 in magnitude:  $|\lambda_i| < 1$ .
- ◊ 9.2.20. The *stable subspace*  $S \subset \mathbb{R}^n$  for a linear iterative system  $\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}$  is defined as the set of all points  $\mathbf{a}$  such that the solution with initial condition  $\mathbf{u}^{(0)} = \mathbf{a}$  satisfies  $\mathbf{u}^{(k)} \rightarrow \mathbf{0}$  as  $k \rightarrow \infty$ . (a) Prove that  $S$  is an invariant subspace for the matrix  $T$ . (b) Determine necessary and sufficient conditions for  $\mathbf{a} \in S$ . (c) Find the stable subspace for the linear systems in Exercise 9.1.14

- ◇ 9.2.21. Consider a second order iterative system  $\mathbf{u}^{(k+2)} = A\mathbf{u}^{(k+1)} + B\mathbf{u}^{(k)}$ , where  $A, B$  are  $n \times n$  matrices. Define a *quadratic eigenvalue* to be a complex number that satisfies  $\det(\lambda^2 \mathbf{I} - \lambda A - B) = 0$ . Prove that the zero solution is globally asymptotically stable if and only if all its quadratic eigenvalues satisfy  $|\lambda| < 1$ .
- ◇ 9.2.22. Let  $p(t)$  be a polynomial. Assume  $0 < \lambda < \mu$ . Prove that there is a positive constant  $C$  such that  $p(n) \lambda^n < C \mu^n$  for all  $n > 0$ .

## Fixed Points

The zero vector  $\mathbf{0}$  is always a *fixed point* for a linear iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$ , since  $\mathbf{0} = T\mathbf{0}$ , and so  $\mathbf{u}^{(k)} \equiv \mathbf{0}$  is an equilibrium solution. Are there any others? The answer is immediate:  $\mathbf{u}^*$  is a fixed point if and only if  $\mathbf{u}^* = T\mathbf{u}^*$ , and hence  $\mathbf{u}^*$  satisfies the eigenvalue equation for  $T$  with for the unit eigenvalue  $\lambda = 1$ . Thus, the system admits a nonzero fixed point if and only if the coefficient matrix  $T$  has 1 as an eigenvalue. Since every nonzero scalar multiple of the eigenvector  $\mathbf{u}^*$  is also an eigenvector, in such cases the system has infinitely many fixed points, namely all elements of the eigenspace  $V_1 = \ker(T - \mathbf{I})$ , including  $\mathbf{0}$ . We are interested in whether the fixed points are *stable* in the sense that solutions having nearby initial conditions remain nearby. More precisely:

**Definition 9.16.** A fixed point  $\mathbf{u}^*$  of an iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$  is called *stable* if for every  $\varepsilon > 0$  there exists a  $\delta > 0$  such that whenever  $\|\mathbf{u}^{(0)} - \mathbf{u}^*\| < \delta$ , then the resulting iterates satisfy  $\|\mathbf{u}^{(k)} - \mathbf{u}^*\| < \varepsilon$  for all  $k$ .

The stability of the fixed points, at least if the coefficient matrix is complete, is governed by the same solution formula (9.9). If the eigenvalue  $\lambda_1 = 1$  is simple, and all other eigenvalues are less than one in modulus, so

$$1 = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_n|,$$

then the solution takes the asymptotic form

$$\mathbf{u}^{(k)} = c_1 \mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2 + \dots + c_n \lambda_n^k \mathbf{v}_n \longrightarrow c_1 \mathbf{v}_1, \quad \text{as } k \rightarrow \infty, \quad (9.29)$$

converging to one of the fixed points, i.e., to a multiple of the eigenvector  $\mathbf{v}_1$ . The coefficient  $c_1$  is prescribed by the initial conditions, cf. (9.10). The rate of convergence of the solution is governed by the modulus  $|\lambda_2|$  of the *subdominant eigenvalue*.

**Proposition 9.17.** Suppose that  $T$  has a simple (or, more generally, complete) eigenvalue  $\lambda_1 = 1$ , and, moreover, all other eigenvalues satisfy  $|\lambda_j| < 1$ . Then all solutions to the linear iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$  converge to a vector  $\mathbf{v} \in V_1$  that lies in the  $\lambda_1 = 1$  eigenspace. Moreover, all the fixed points  $\mathbf{v} \in V_1$  of  $T$  are *stable*.

Stability of a fixed point does not imply asymptotic stability, since nearby solutions may converge to a nearby fixed point, i.e., a slightly different element of the eigenspace  $V_1$ .

The general necessary and sufficient conditions for stability of the fixed points of a linear iterative system is governed by the spectral radius of its coefficient matrix, as follows. The proof is relegated to Exercise 9.2.28.

**Theorem 9.18.** The fixed points of an iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$  are stable if and only if  $\rho(T) \leq 1$  and, moreover, every eigenvalue of modulus  $|\lambda| = 1$  is complete.

Thus, with regard to linear iterative systems, either all fixed points are stable or all are unstable. Keep in mind that the fixed points are the elements of the eigenspace  $V_1$  corresponding to the eigenvalue  $\lambda = 1$ , if such exists. If 1 is not an eigenvalue of  $T$ , then  $\mathbf{u}^* = \mathbf{0}$  is the only fixed point.

**Example 9.19.** Consider the iterative system with coefficient matrix

$$T = \begin{pmatrix} \frac{3}{2} & -\frac{1}{2} & -3 \\ -\frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

The eigenvalues and corresponding eigenvectors are

$$\begin{aligned} \lambda_1 &= 1, & \lambda_2 &= \frac{1+i}{2}, & \lambda_3 &= \frac{1-i}{2}, \\ \mathbf{v}_1 &= \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 2-i \\ -1 \\ 1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 2+i \\ -1 \\ 1 \end{pmatrix}. \end{aligned}$$

Since  $\lambda_1 = 1$ , every scalar multiple of the eigenvector  $\mathbf{v}_1$  is a fixed point. The fixed points are stable, since the remaining eigenvalues have modulus  $|\lambda_2| = |\lambda_3| = \frac{1}{2}\sqrt{2} \approx .7071 < 1$ . Thus, the iterates  $\mathbf{u}^{(k)} = T^k \mathbf{a} \rightarrow c_1 \mathbf{v}_1$  will eventually converge to a multiple of the first eigenvector; in almost all cases the convergence rate is  $\frac{1}{2}\sqrt{2}$ . For example, starting with  $\mathbf{u}^{(0)} = (1, 1, 1)^T$ , leads to the iterates<sup>†</sup>

$$\begin{aligned} \mathbf{u}^{(5)} &= \begin{pmatrix} -9.5 \\ 4.75 \\ -2.75 \end{pmatrix}, & \mathbf{u}^{(10)} &= \begin{pmatrix} -7.9062 \\ 3.9062 \\ -1.9062 \end{pmatrix}, & \mathbf{u}^{(15)} &= \begin{pmatrix} -7.9766 \\ 4.0 \\ -2.0 \end{pmatrix}, \\ \mathbf{u}^{(20)} &= \begin{pmatrix} -8.0088 \\ 4.0029 \\ -2.0029 \end{pmatrix}, & \mathbf{u}^{(25)} &= \begin{pmatrix} -7.9985 \\ 3.9993 \\ -1.9993 \end{pmatrix}, & \mathbf{u}^{(30)} &= \begin{pmatrix} -8.0001 \\ 4.0001 \\ -2.0001 \end{pmatrix}, \end{aligned}$$

which are gradually converging to the particular eigenvector  $(-8, 4, -2)^T = -2\mathbf{v}_1$ . This can be predicted in advance by decomposing the initial vector into a linear combination of the eigenvectors:

$$\mathbf{u}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -2 \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + \frac{3+3i}{2} \begin{pmatrix} 2-i \\ -1 \\ 1 \end{pmatrix} + \frac{3-3i}{2} \begin{pmatrix} 2+i \\ -1 \\ 1 \end{pmatrix},$$

whence

$$\mathbf{u}^{(k)} = \begin{pmatrix} -8 \\ 4 \\ -2 \end{pmatrix} + \frac{3+3i}{2} \left( \frac{1+i}{2} \right)^k \begin{pmatrix} 2-i \\ -1 \\ 1 \end{pmatrix} + \frac{3-3i}{2} \left( \frac{1-i}{2} \right)^k \begin{pmatrix} 2+i \\ -1 \\ 1 \end{pmatrix},$$

and so  $\mathbf{u}^{(k)} \rightarrow (-8, 4, -2)^T$  as  $k \rightarrow \infty$ . Despite the complex formula, the solution is, in fact, real.

---

<sup>†</sup> Since the convergence is slow, we only display every fifth one.

## Exercises

9.2.23. Find all fixed points for the iterative systems with the following coefficient matrices:

$$(a) \begin{pmatrix} .7 & .3 \\ .2 & .8 \end{pmatrix}, \quad (b) \begin{pmatrix} .6 & 1.0 \\ .3 & -.7 \end{pmatrix}, \quad (c) \begin{pmatrix} -1 & -1 & -4 \\ -2 & 0 & -4 \\ 1 & -1 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 2 & 1 & -1 \\ 2 & 3 & -2 \\ -1 & -1 & 2 \end{pmatrix}.$$

9.2.24. Discuss the stability of each fixed point and the asymptotic behavior(s) of the solutions to the systems in Exercise 9.2.23. Which fixed point, if any, does the solution with initial condition  $\mathbf{u}^{(0)} = \mathbf{e}_1$  converge to?

9.2.25. Suppose  $T$  is a symmetric matrix that satisfies the hypotheses of Proposition 9.17 with a simple eigenvalue  $\lambda_1 = 1$ . Prove that the solution  $\mathbf{u}^{(k)}$  to the linear iterative system

$$\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} \text{ has limiting value } \lim_{k \rightarrow \infty} \mathbf{u}^{(k)} = \frac{\mathbf{u}^{(0)} \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} \mathbf{v}_1.$$

9.2.26. *True or false:* If  $T$  has a stable nonzero fixed point, then it is a convergent matrix.

9.2.27. *True or false:* If every point  $\mathbf{u} \in \mathbb{R}^n$  is a fixed point, then they are all stable. Can you characterize such systems?

◇ 9.2.28. Prove Theorem 9.18: (a) assuming  $T$  is complete, (b) for general  $T$ .

*Hint:* Use Exercise 9.1.40.

◇ 9.2.29. (a) Under what conditions does the linear iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$  have a *period 2 solution*, meaning that the iterates repeat after every other iterate:  $\mathbf{u}^{(k+2)} = \mathbf{u}^{(k)} \neq \mathbf{u}^{(k+1)}$ ? Give an example of such a system. (b) Under what conditions is there a unique period 2 solution? (c) What about a period  $m$  solution for  $2 < m \in \mathbb{N}$ ?

## Matrix Norms and Convergence

As we now know, the convergence of a linear iterative system is governed by the spectral radius, or, equivalently, the modulus of the largest eigenvalue of the coefficient matrix. Unfortunately, finding accurate approximations to the eigenvalues of most matrices is a nontrivial computational task. Indeed, as we will learn in Section 9.5, all practical numerical algorithms rely on some form of iteration. But using iteration to determine the spectral radius defeats the purpose, which is to predict the behavior of the iterative system in advance! One independent means of accomplishing this is through matrix norms, as introduced at the end of Section 3.3.

Let  $\|\mathbf{v}\|$  denote a norm on<sup>†</sup>  $\mathbb{R}^n$ . Theorem 3.20 defines the induced natural matrix norm on the space of  $n \times n$  matrices, denoted by  $\|A\| = \max\{\|A\mathbf{u}\| \mid \|\mathbf{u}\| = 1\}$ . The following result relates the magnitude of the norm of a matrix to convergence of the associated iterative system.

**Proposition 9.20.** If  $A$  is a square matrix, then  $\|A^k\| \leq \|A\|^k$ . In particular, if  $\|A\| < 1$ , then  $\|A^k\| \rightarrow 0$  as  $k \rightarrow \infty$ , and hence  $A$  is a convergent matrix:  $A^k \rightarrow \mathbf{O}$ .

The first part is a restatement of Proposition 3.22, and the second part is an immediate consequence. The converse to this result is not quite true; a convergent matrix does not

<sup>†</sup> We work with real iterative systems throughout this chapter, but the methods readily extend to their complex counterparts.

necessarily have matrix norm less than 1, or even  $\leq 1$  — see Example 9.23 below. An alternative proof of Proposition 9.20 can be based on the following useful estimate:

**Theorem 9.21.** The spectral radius of a matrix is bounded by its matrix norm:

$$\rho(A) \leq \|A\|. \quad (9.30)$$

*Proof:* If  $\lambda$  is a real eigenvalue, and  $\mathbf{u}$  a corresponding unit eigenvector, so that  $A\mathbf{u} = \lambda \mathbf{u}$  with  $\|\mathbf{u}\| = 1$ , then

$$\|A\mathbf{u}\| = \|\lambda \mathbf{u}\| = |\lambda| \|\mathbf{u}\| = |\lambda|. \quad (9.31)$$

Since  $\|A\|$  is the maximum of  $\|A\mathbf{u}\|$  over all possible unit vectors, this implies that

$$|\lambda| \leq \|A\|. \quad (9.32)$$

If all the eigenvalues of  $A$  are real, then the spectral radius is the maximum of their absolute values, and so it too is bounded by  $\|A\|$ , proving (9.30).

If  $A$  has complex eigenvalues, then we need to work a little harder to establish (9.32). (This is because the matrix norm is defined by the effect of  $A$  on *real* vectors, and so we cannot directly use the complex eigenvectors to establish the required bound.) Let  $\lambda = r e^{i\theta}$  be a complex eigenvalue with complex eigenvector  $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ . Define

$$\mu = \min \{ \|\operatorname{Re}(e^{i\varphi} \mathbf{z})\| = \|(\cos \varphi) \mathbf{x} - (\sin \varphi) \mathbf{y}\| \mid 0 \leq \varphi \leq 2\pi \}. \quad (9.33)$$

Since the indicated subset is a closed curve (in fact, an ellipse) that does not go through the origin<sup>†</sup>,  $\mu > 0$ . Let  $\varphi_0$  denote the value of the angle that produces the minimum, so

$$\mu = \|(\cos \varphi_0) \mathbf{x} - (\sin \varphi_0) \mathbf{y}\| = \|\operatorname{Re}(e^{i\varphi_0} \mathbf{z})\|.$$

Define the real unit vector

$$\mathbf{u} = \frac{\operatorname{Re}(e^{i\varphi_0} \mathbf{z})}{\mu} = \frac{(\cos \varphi_0) \mathbf{x} - (\sin \varphi_0) \mathbf{y}}{\mu}, \quad \text{so that} \quad \|\mathbf{u}\| = 1.$$

Then

$$A\mathbf{u} = \frac{1}{\mu} \operatorname{Re}(e^{i\varphi_0} A \mathbf{z}) = \frac{1}{\mu} \operatorname{Re}(e^{i\varphi_0} r e^{i\theta} \mathbf{z}) = \frac{r}{\mu} \operatorname{Re}(e^{i(\varphi_0+\theta)} \mathbf{z}).$$

Therefore, keeping in mind that  $m$  is the minimal value in (9.33),

$$\|A\| \geq \|A\mathbf{u}\| = \frac{r}{\mu} \|\operatorname{Re}(e^{i(\varphi_0+\theta)} \mathbf{z})\| \geq r = |\lambda|, \quad (9.34)$$

and so (9.32) also holds for complex eigenvalues. Q.E.D.

Let us see what the convergence criterion of Proposition 9.20 says for a couple of our well-known matrix norms. First, the formula (3.44) for the  $\infty$  norm implies the following convergence criterion.

**Proposition 9.22.** If all the absolute row sums of  $A$  are strictly less than 1, then  $\|A\|_\infty < 1$  and hence  $A$  is a convergent matrix.

**Example 9.23.** Consider the symmetric matrix  $A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{4} \end{pmatrix}$ . Its two absolute row sums are  $\left| \frac{1}{2} \right| + \left| -\frac{1}{3} \right| = \frac{5}{6}$ ,  $\left| -\frac{1}{3} \right| + \left| \frac{1}{4} \right| = \frac{7}{12}$ , so

$$\|A\|_\infty = \max \left\{ \frac{5}{6}, \frac{7}{12} \right\} = \frac{5}{6} = .83333 \dots$$

<sup>†</sup> This relies on the fact that  $\mathbf{x}, \mathbf{y}$  are linearly independent, which was shown in Exercise 8.3.12.

Since the norm is less than 1,  $A$  is a convergent matrix. Indeed, its eigenvalues are

$$\lambda_1 = \frac{9 + \sqrt{73}}{24} = .731000\dots, \quad \lambda_2 = \frac{9 - \sqrt{73}}{24} = .018999\dots,$$

and hence the spectral radius is  $\rho(A) = \lambda_1 = .731000\dots$ , which is slightly smaller than its  $\infty$  norm.

The row sum test for convergence is not always conclusive. For example, the matrix

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{3}{5} \\ -\frac{3}{5} & \frac{1}{4} \end{pmatrix} \quad \text{has matrix norm} \quad \|A\|_\infty = \frac{11}{10} > 1.$$

On the other hand, its eigenvalues are  $\frac{15 \pm \sqrt{601}}{40}$  and hence its spectral radius is  $\rho(A) = \frac{15 + \sqrt{601}}{40} = .987882\dots$ , which implies that  $A$  is (just barely) convergent, even though its maximal row sum is larger than 1.

Similarly, using the formula (8.61) for the Euclidean matrix norm, one deduces a convergence criterion based on the magnitude of the singular values.

**Proposition 9.24.** If  $A$  is a square matrix whose largest singular value satisfies  $\sigma_1 < 1$ , then  $\|A\|_2 < 1$  and hence  $A$  is a convergent matrix.

**Example 9.25.** Consider the matrix and associated Gram matrix

$$A = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}, \quad A^T A = \begin{pmatrix} .2225 & .0800 & .1250 \\ .0800 & .1511 & -.1111 \\ .1250 & -.1111 & .3611 \end{pmatrix}.$$

Then  $A^T A$  has eigenvalues  $\lambda_1 = .4472$ ,  $\lambda_2 = .2665$ ,  $\lambda_3 = .0210$ , and hence the singular values of  $A$  are their square roots:  $\sigma_1 = .6687$ ,  $\sigma_2 = .5163$ ,  $\sigma_3 = .1448$ . The Euclidean matrix norm of  $A$  is the largest singular value, and so  $\|A\|_2 = .6687$ , proving that  $A$  is a convergent matrix. Note that, as always, the matrix norm overestimates the spectral radius, which is  $\rho(A) = .5$ .

Unfortunately, as we discovered in Example 9.23, matrix norms are not a foolproof test of convergence. There exist convergent matrices such that  $\rho(A) < 1$  that yet have matrix norm  $\|A\| \geq 1$ . In such cases, the matrix norm is not able to predict convergence of the iterative system, although one should expect the convergence to be quite slow. Although such pathology might show up in the chosen matrix norm, it turns out that one can always rig up some matrix norm for which  $\|A\| < 1$ . This follows from a more general result, whose proof can be found in [62].

**Theorem 9.26.** Let  $A$  have spectral radius  $\rho(A)$ . If  $\varepsilon > 0$  is any positive number, then there exists a matrix norm  $\|\cdot\|$  such that

$$\rho(A) \leq \|A\| < \rho(A) + \varepsilon. \tag{9.35}$$

**Corollary 9.27.** If  $A$  is a convergent matrix, then there exists a matrix norm such that  $\|A\| < 1$ .

*Proof:* By definition,  $A$  is convergent if and only if  $\rho(A) < 1$ . Choose  $\varepsilon > 0$  such that  $\rho(A) + \varepsilon < 1$ . Any norm that then satisfies (9.35) has the desired property. *Q.E.D.*

It can also be proved, [48], that, given a matrix norm,  $\lim_{n \rightarrow \infty} \|A^n\|^{1/n} = \rho(A)$ , and hence, if  $A$  is convergent, then  $\|A^n\| < 1$  for  $n$  sufficiently large.

**Warning.** Based on the accumulated evidence, one might be tempted to speculate that the spectral radius itself defines a matrix norm. Unfortunately, this is not the case. For example, the nonzero matrix  $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$  has zero spectral radius,  $\rho(A) = 0$ , in violation of a basic norm axiom.

## Exercises

9.2.30. Compute the  $\infty$  matrix norm of the following matrices. Which are guaranteed to be

convergent? (a)  $\begin{pmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{6} \end{pmatrix}$ , (b)  $\begin{pmatrix} \frac{5}{3} & \frac{4}{3} \\ -\frac{7}{6} & -\frac{5}{6} \end{pmatrix}$ , (c)  $\begin{pmatrix} \frac{2}{7} & -\frac{2}{7} \\ -\frac{2}{7} & \frac{6}{7} \end{pmatrix}$ , (d)  $\begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ -\frac{1}{2} & \frac{5}{4} \end{pmatrix}$ ,  
 (e)  $\begin{pmatrix} \frac{2}{7} & \frac{2}{7} & -\frac{4}{7} \\ 0 & \frac{2}{7} & \frac{6}{7} \\ \frac{2}{7} & \frac{4}{7} & \frac{2}{7} \end{pmatrix}$ , (f)  $\begin{pmatrix} 0 & .1 & .8 \\ -.1 & 0 & .1 \\ -.8 & -.1 & 0 \end{pmatrix}$ , (g)  $\begin{pmatrix} 1 & -\frac{2}{3} & -\frac{2}{3} \\ 1 & -\frac{1}{3} & -1 \\ \frac{1}{3} & -\frac{2}{3} & 0 \end{pmatrix}$ , (h)  $\begin{pmatrix} \frac{1}{3} & 0 & 0 \\ -\frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{pmatrix}$ .

9.2.31. Compute the Euclidean matrix norm of each matrix in Exercise 9.2.30. Have your convergence conclusions changed?

9.2.32. Compute the spectral radii of the matrices in Exercise 9.2.30. Which are convergent? Compare your conclusions with those of Exercises 9.2.30 and 9.2.31.

9.2.33. Let  $k$  be an integer and set  $A_k = \begin{pmatrix} k & -1 \\ k^2 & -k \end{pmatrix}$ . Compute (a)  $\|A_k\|_\infty$ , (b)  $\|A_k\|_2$ , (c)  $\rho(A_k)$ . (d) Explain why every  $A_k$  is a convergent matrix, even though their matrix norms can be arbitrarily large. (e) Why does this not contradict Corollary 9.27?

9.2.34. Show that if  $|c| < 1/\|A\|$ , then  $cA$  is a convergent matrix.

◊ 9.2.35. Prove that the spectral radius function does *not* satisfy the triangle inequality by finding matrices  $A, B$  such that  $\rho(A + B) > \rho(A) + \rho(B)$ .

9.2.36. Find a convergent matrix that has dominant singular value  $\sigma_1 > 1$ .

◊ 9.2.37. Prove that if  $A$  is a real symmetric matrix, then its Euclidean matrix norm is equal to its spectral radius.

◊ 9.2.38. Let  $A$  be a square matrix. Let  $s = \max\{s_1, \dots, s_n\}$  be the maximal absolute row sum of  $A$  and let  $t = \min\{|a_{ii}| - r_i\}$ , with  $r_i$  given by (8.27). Prove that  $\max\{0, t\} \leq \rho(A) \leq s$ .

9.2.39. Suppose the largest entry (in modulus) of  $A$  is  $|a_{ij}| = a_*$ . Can you bound its radius of convergence?

9.2.40. (a) Suppose that every entry of the  $n \times n$  matrix  $A$  is bounded by  $|a_{ij}| < 1/n$ . Prove that  $A$  is a convergent matrix. *Hint:* Use Exercise 9.2.38. (b) Produce a matrix of size  $n \times n$  with one or more entries satisfying  $|a_{ij}| = 1/n$  that is not convergent.

9.2.41. Write down an example of a strictly diagonally dominant matrix that is also convergent.

9.2.42. *True or false:* If  $B = S^{-1}AS$  are similar matrices, then

(a)  $\|B\|_\infty = \|A\|_\infty$ , (b)  $\|B\|_2 = \|A\|_2$ , (c)  $\rho(B) = \rho(A)$ .

9.2.43. Prove that the curve parametrized in (9.33) is an ellipse. What are its semi-axes?

- ◇ 9.2.44. (a) Prove that the individual entries  $a_{ij}$  of a matrix  $A$  are bounded in absolute value by its  $\infty$  matrix norm:  $|a_{ij}| \leq \|A\|_\infty$ . (b) Prove that if the series  $\sum_{n=0}^{\infty} \|A_n\|_\infty < \infty$  converges, then the *matrix series*  $\sum_{n=0}^{\infty} A_n = A^*$  converges to some matrix  $A^*$ .  
 (c) Let  $\|A\|$  denote any natural matrix norm. Prove that if the series  $\sum_{n=0}^{\infty} \|A_n\| < \infty$  converges, then the matrix series  $\sum_{n=0}^{\infty} A_n = A^*$  converges.
- 9.2.45. (a) Use Exercise 9.2.44 to prove that the *geometric matrix series*  $\sum_{n=0}^{\infty} A^n$  converges whenever  $\rho(A) < 1$ . *Hint:* Apply Corollary 9.27.  
 (b) Prove that the sum equals  $(I - A)^{-1}$ . How do you know  $I - A$  is invertible?

### 9.3 Markov Processes

A discrete probabilistic process in which the future state of a system depends only upon its current configuration is known as a *Markov chain*, to honor the pioneering early twentieth studies of the Russian mathematician Andrei Markov. Markov chains are described by linear iterative systems whose coefficient matrices have a special form. They define the simplest examples of stochastic processes, [4, 23], which have many profound physical, biological, economic, and statistical applications, including networks, internet search engines, speech recognition, and routing.

To take a very simple (albeit slightly artificial) example, suppose you would like to be able to predict the weather in your city. Consulting local weather records over the past decade, you determine that

- (a) If today is sunny, there is a 70% chance that tomorrow will also be sunny,
- (b) But, if today is cloudy, the chances are 80% that tomorrow will also be cloudy.

Question: given that today is sunny, what is the probability that next Saturday's weather will also be sunny?

To formulate this process mathematically, we let  $s^{(k)}$  denote the probability that day  $k$  is sunny and  $c^{(k)}$  the probability that it is cloudy. If we assume that these are the only possibilities, then the individual probabilities must sum to 1, so

$$s^{(k)} + c^{(k)} = 1.$$

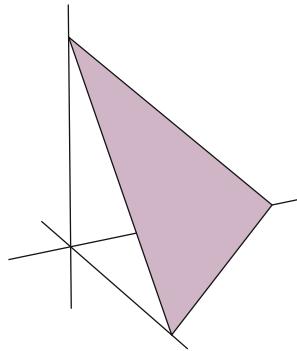
According to our data, the probability that the next day is sunny or cloudy is expressed by the equations

$$s^{(k+1)} = .7 s^{(k)} + .2 c^{(k)}, \quad c^{(k+1)} = .3 s^{(k)} + .8 c^{(k)}. \quad (9.36)$$

Indeed, day  $k + 1$  could be sunny either if day  $k$  was, with a 70% chance, or, if day  $k$  was cloudy, there is still a 20% chance of day  $k + 1$  being sunny. We rewrite (9.36) in a more convenient matrix form:

$$\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad \text{where} \quad T = \begin{pmatrix} .7 & .2 \\ .3 & .8 \end{pmatrix}, \quad \mathbf{u}^{(k)} = \begin{pmatrix} s^{(k)} \\ c^{(k)} \end{pmatrix}. \quad (9.37)$$

In a Markov process, the vector of probabilities  $\mathbf{u}^{(k)}$  is known as the  $k^{\text{th}}$  *state vector* and the matrix  $T$  is known as the *transition matrix*, whose entries fix the transition probabilities between the various states.



**Figure 9.4.** The Set of Probability Vectors in  $\mathbb{R}^3$ .

By assumption, the initial state vector is  $\mathbf{u}^{(0)} = (1, 0)^T$ , since we know for certain that today is sunny. Rounded off to three decimal places, the subsequent state vectors are

$$\begin{aligned}\mathbf{u}^{(1)} &\simeq \begin{pmatrix} .7 \\ .3 \end{pmatrix}, & \mathbf{u}^{(2)} &\simeq \begin{pmatrix} .55 \\ .45 \end{pmatrix}, & \mathbf{u}^{(3)} &\simeq \begin{pmatrix} .475 \\ .525 \end{pmatrix}, & \mathbf{u}^{(4)} &\simeq \begin{pmatrix} .438 \\ .563 \end{pmatrix}, \\ \mathbf{u}^{(5)} &\simeq \begin{pmatrix} .419 \\ .581 \end{pmatrix}, & \mathbf{u}^{(6)} &\simeq \begin{pmatrix} .410 \\ .591 \end{pmatrix}, & \mathbf{u}^{(7)} &\simeq \begin{pmatrix} .405 \\ .595 \end{pmatrix}, & \mathbf{u}^{(8)} &\simeq \begin{pmatrix} .402 \\ .598 \end{pmatrix}.\end{aligned}$$

The iterates converge fairly rapidly to  $(.4, .6)^T$ , which is, in fact, a fixed point for the iterative system (9.37). Thus, in the long run, 40% of the days will be sunny and 60% will be cloudy. Let us explain why this happens.

**Definition 9.28.** A vector  $\mathbf{u} = (u_1, u_2, \dots, u_n)^T \in \mathbb{R}^n$  is called a *probability vector* if all its entries lie between 0 and 1, so  $0 \leq u_i \leq 1$  for  $i = 1, \dots, n$ , and, moreover, their sum is  $u_1 + \dots + u_n = 1$ .

We interpret the entry  $u_i$  of a probability vector as the probability that the system is in state number  $i$ . The fact that the entries add up to 1 means that they represent a complete list of probabilities for the possible states of the system. The set of probability vectors defines an  $(n-1)$ -dimensional *simplex* in  $\mathbb{R}^n$ . For example, the possible probability vectors  $\mathbf{u} \in \mathbb{R}^3$  fill the equilateral triangle plotted in Figure 9.4.

**Remark.** Every nonzero vector  $\mathbf{0} \neq \mathbf{v} = (v_1, v_2, \dots, v_n)^T$  with all non-negative entries,  $v_i \geq 0$  for  $i = 1, \dots, n$ , can be converted into a parallel probability vector by dividing by the sum of its entries:

$$\mathbf{u} = \frac{\mathbf{v}}{v_1 + \dots + v_n}. \quad (9.38)$$

For example, if  $\mathbf{v} = (3, 2, 0, 1)^T$ , then  $\mathbf{u} = (\frac{1}{2}, \frac{1}{3}, 0, \frac{1}{6})^T$  is the corresponding probability vector.

In general, a *Markov chain* is represented by a first order linear iterative system

$$\mathbf{u}^{(k+1)} = T \mathbf{u}^{(k)}, \quad (9.39)$$

whose initial state  $\mathbf{u}^{(0)}$  is a probability vector. The entries of the *transition matrix*  $T$  must satisfy

$$0 \leq t_{ij} \leq 1, \quad t_{1j} + \dots + t_{nj} = 1. \quad (9.40)$$

The entry  $t_{ij}$  represents the *transitional probability* that the system will switch from state  $j$  to state  $i$ . (Note the reversal of indices.) Since this covers all possible transitions, the *column sums* of the transition matrix are all equal to 1, and hence each column of  $T$  is a probability vector, which is equivalent to condition (9.40). In Exercise 9.3.24 you are asked to show that, under these assumptions, if  $\mathbf{u}^{(k)}$  is a probability vector, then so is  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)}$ , and hence, given our assumption on the initial state, the solution  $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)}$  to the Markov process defines a sequence, or “chain”, of probability vectors.

Let us now investigate the convergence of the Markov chain. Not all Markov chains converge — see Exercise 9.3.9 for an example — and so we impose some additional mild restrictions on the transition matrix.

**Definition 9.29.** A transition matrix (9.40) is *regular* if some power  $T^k$  contains no zero entries. In particular, if  $T$  itself has no zero entries, then it is regular.

**Warning.** The term “regular transition matrix” has nothing to do with our earlier term “regular matrix”, which was used to describe matrices with an *LU* factorization.

The entries of  $T^k$  describe the transition probabilities of getting from one state to another in  $k$  steps. Thus, regularity of the transition matrix means that there is a nonzero probability of getting from any state to any other state in exactly  $k$  steps for some  $k \geq 1$ .

The asymptotic behavior of a regular Markov chain is governed by the following basic result, originally due to the German mathematicians Oskar Perron and Georg Frobenius in the early part of the twentieth century. A proof can be found at the end of this section.

**Theorem 9.30.** If  $T$  is a regular transition matrix, then it admits a unique *probability eigenvector*  $\mathbf{u}^*$  with eigenvalue  $\lambda_1 = 1$ . Moreover, a Markov chain with coefficient matrix  $T$  will converge to the probability eigenvector:  $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^*$  as  $k \rightarrow \infty$ .

**Example 9.31.** The eigenvalues and eigenvectors of the weather transition matrix (9.37) are

$$\lambda_1 = 1, \quad \mathbf{v}_1 = \begin{pmatrix} \frac{2}{3} \\ 1 \end{pmatrix}, \quad \lambda_2 = .5, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The first eigenvector is then converted into a probability vector via formula (9.38):

$$\mathbf{u}^* = \mathbf{u}_1 = \frac{1}{1 + \frac{2}{3}} \begin{pmatrix} \frac{2}{3} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \end{pmatrix}.$$

This distinguished probability eigenvector represents the final asymptotic state of the system after many iterations, *no matter what the initial state is*. Thus, our earlier observation that about 40% of the days will be sunny and 60% will be cloudy does not depend upon today’s weather.

**Example 9.32.** A taxi company in Minnesota serves the cities of Minneapolis and St. Paul, as well as the nearby suburbs. Records indicate that, on average, 10% of the customers taking a taxi in Minneapolis go to St. Paul and 30% go to the suburbs. Customers boarding in St. Paul have a 30% chance of going to Minneapolis and a 30% chance of going to the suburbs, while suburban customers choose Minneapolis 40% of the time and St. Paul 30% of the time. The owner of the taxi company is interested in knowing where the taxis will end up, on average. Let us write this as a Markov process. The entries of the state vector  $\mathbf{u}^{(k)} = (u_1^{(k)}, u_2^{(k)}, u_3^{(k)})^T$  tell what proportion of the taxi fleet is, respec-

tively, in Minneapolis, St. Paul, and the suburbs, or, equivalently, the probability that an individual taxi will be in one of the three locations. Using the given data, we construct the relevant transition matrix

$$T = \begin{pmatrix} .6 & .3 & .4 \\ .1 & .4 & .3 \\ .3 & .3 & .3 \end{pmatrix}.$$

Note that  $T$  is regular since it has no zero entries. The probability eigenvector

$$\mathbf{u}^* \simeq (.4714, .2286, .3)^T$$

corresponding to the unit eigenvalue  $\lambda_1 = 1$  is found by first solving the linear system  $(T - I)\mathbf{v}^* = 0$  and then converting the solution<sup>†</sup>  $\mathbf{v}^*$  into a valid probability vector  $\mathbf{u}^*$  by use of formula (9.38). According to Theorem 9.30, no matter how the taxis are initially distributed, eventually about 47% of the taxis will be in Minneapolis, 23% in St. Paul, and 30% in the suburbs. This can be confirmed by running numerical experiments. Moreover, if the owner places this fraction of the taxis in the three locations, then they will more or less remain in such proportions forever.

**Remark.** As noted earlier — see Proposition 9.17 — the convergence rate of the Markov chain to its steady state is governed by the size of the *subdominant eigenvalue*  $\lambda_2$ . The smaller  $|\lambda_2|$  is, the faster the process converges. In the taxi example,  $\lambda_2 = .3$  (and  $\lambda_3 = 0$ ), and so the convergence to steady state is fairly rapid.

A Markov process can also be viewed as a weighted digraph. Each state corresponds to a vertex. A nonzero transition probability from one state to another corresponds to a weighted directed edge between the two vertices. Note that the digraph is typically not simple, since vertices can have two edges connecting them, one representing the transition probability of getting from the first to the second, and the second edge representing the transition probability of going in the other direction. The original PageRank algorithm that underlies Google's search engine, [64, 52], starts with the internet digraph, whose vertices are web pages and whose directed edges represent links from one web page to another, which are weighted according to the number of such links. To be effective, the resulting weighted internet digraph is supplemented by adding in a number of random low weight edges. One then computes the probability eigenvector associated with the resulting digraph-based Markov process, the magnitudes of whose entries, indexed by the nodes, effectively rank the corresponding web pages.

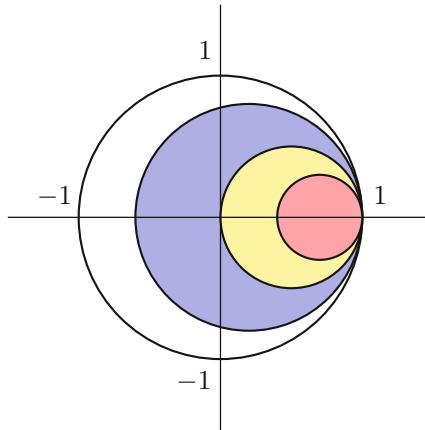
*Proof of Theorem 9.30:* We begin the proof by replacing  $T$  by its transpose<sup>‡</sup>  $M = T^T$ , keeping in mind that every eigenvalue of  $T$  is also an eigenvalue of  $M$  albeit with different eigenvectors, cf. Proposition 8.12. The conditions (9.40) tell us that the matrix  $M$  has entries  $0 \leq m_{ij} = t_{ji} \leq 1$ , and, moreover, the *row sums*  $s_i = \sum_{j=1}^n m_{ij} = 1$  of  $M$ , being the same as the corresponding column sums of  $T$ , are all equal to 1. Since  $M^k = (T^k)^T$ , regularity of  $T$  implies that some power  $M^k$  has all positive entries.

According to Exercise 1.2.29, if  $\mathbf{z} = (1, \dots, 1)^T$  is the column vector all of whose entries are equal to 1, then the entries of  $M\mathbf{z}$  are the row sums of  $M$ . Therefore,  $M\mathbf{z} = \mathbf{z}$ , which implies that  $\mathbf{z}$  is an eigenvector of  $M$  with eigenvalue  $\lambda_1 = 1$ . As a consequence,  $T$  also has

---

<sup>†</sup> Theorem 9.30 guarantees that there is an eigenvector  $\mathbf{v}$  with all non-negative entries.

<sup>‡</sup> We apologize for the unfortunate clash of notation when writing the transpose of the matrix  $T$ .



**Figure 9.5.** Gershgorin Disks for a Regular Transition Matrix.

1 as an eigenvalue. Observe that  $\mathbf{z}$  is *not* in general an eigenvector of  $T$ ; indeed, it satisfies the *co-eigenvector equation*  $M\mathbf{z} = T^T\mathbf{z} = \mathbf{z}$ .

We claim that  $\lambda_1 = 1$  is a simple eigenvalue. To this end, we prove that  $\mathbf{z}$  spans the one-dimensional eigenspace  $V_1$ . In other words, we need to show that if  $M\mathbf{v} = \mathbf{v}$ , then its entries  $v_1 = \dots = v_n = a$  are all equal, and so  $\mathbf{v} = a\mathbf{z}$  is a scalar multiple of the known eigenvector  $\mathbf{z}$ . Let us first prove this assuming that all of the entries of  $M$  are strictly positive, and so  $0 < m_{ij} = t_{ji} < 1$  for all  $i, j$ . Suppose  $\mathbf{v}$  is an eigenvector with not all equal entries. Let  $v_k$  be the minimal entry of  $\mathbf{v}$ , so  $v_k \leq v_i$  for all  $i \neq k$ , and at least one inequality is strict, say  $v_k < v_j$ . Then the  $k^{\text{th}}$  entry of the eigenvector equation  $\mathbf{v} = M\mathbf{v}$  is

$$v_k = \sum_{j=1}^n m_{kj} v_j > \left( \sum_{j=1}^n m_{kj} \right) v_k = v_k,$$

where the strict inequality follows from the assumed positivity of the entries of  $M$ , and the final equality follows from the fact that  $M$  has unit row sums. Thus, we are led to a contradiction, and the claim follows. If  $M$  has one or more 0 entries, but  $M^k$  has all positive entries, then we apply the previous argument to the equation  $M^k\mathbf{v} = \mathbf{v}$  which follows from  $M\mathbf{v} = \mathbf{v}$ . If  $\lambda_1 = 1$  is a complete eigenvalue, then we are finished. The proof that this is indeed the case is a bit technical, and we refer the reader to [4] for the details.

Finally, let us prove that all the other eigenvalues of  $M$  are less than 1 in modulus. For this we appeal to the Gershgorin Circle Theorem 8.16. Suppose  $M^k$  has all positive entries, denoted by  $m_{ij}^{(k)} > 0$ . Its Gershgorin disk  $D_i$  is centered at  $m_{ii}^{(k)} > 0$  and has radius  $r_i = 1 - m_{ii}^{(k)} < 1$  since the  $i^{\text{th}}$  row sum of  $M^k$  equals 1. Thus the disk lies strictly inside the open unit disk  $|z| < 1$  *except* for a single boundary point at  $z = 1$ ; see Figure 9.5. The Circle Theorem 8.16 implies that all eigenvalues of  $M^k$  except the unit eigenvalue  $\lambda_1 = 1$  must lie strictly inside the unit disk. Since these are just the  $k^{\text{th}}$  powers of the eigenvalues of  $M$ , the same holds for the eigenvalues themselves, so  $|\lambda_j| < 1$  for  $j \geq 2$ .

Therefore, the matrix  $M$ , and, hence, also  $T$ , satisfies the hypotheses of Proposition 9.17. We conclude that the iterates  $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)} \rightarrow \mathbf{u}^*$  converge to a multiple of the probability eigenvector of  $T$ . If the initial condition  $\mathbf{u}^{(0)}$  is a probability vector, then so is every subsequent state vector  $\mathbf{u}^{(k)}$ , and so their limit  $\mathbf{u}^*$  must also be a probability vector. This completes the proof of the theorem. *Q.E.D.*

## Exercises

9.3.1. Determine if the following matrices are regular transition matrices. If so, find the

- associated probability eigenvector. (a)  $\begin{pmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{3}{4} & \frac{2}{3} \end{pmatrix}$ , (b)  $\begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$ , (c)  $\begin{pmatrix} \frac{1}{4} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{3} \end{pmatrix}$ ,  
 (d)  $\begin{pmatrix} 0 & \frac{1}{5} \\ 1 & \frac{4}{5} \end{pmatrix}$ , (e)  $\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , (f)  $\begin{pmatrix} .3 & .5 & .2 \\ .3 & .2 & .5 \\ .4 & .3 & .3 \end{pmatrix}$ , (g)  $\begin{pmatrix} .1 & .5 & .4 \\ .6 & .1 & .3 \\ .3 & 0 & .7 \end{pmatrix}$ ,  
 (h)  $\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$ , (i)  $\begin{pmatrix} 0 & .2 & 0 & 1 \\ .5 & 0 & .3 & 0 \\ 0 & .8 & 0 & 0 \\ .5 & 0 & .7 & 0 \end{pmatrix}$ , (j)  $\begin{pmatrix} .1 & .2 & .3 & .4 \\ .2 & .5 & .3 & .1 \\ .3 & .3 & .1 & .3 \\ .4 & .1 & .3 & .2 \end{pmatrix}$ , (k)  $\begin{pmatrix} 0 & .6 & 0 & .4 \\ .5 & 0 & .3 & .1 \\ 0 & .4 & 0 & .5 \\ .5 & 0 & .7 & 0 \end{pmatrix}$ .

9.3.2. A business executive is managing three branches, labeled  $A$ ,  $B$ , and  $C$ , of a corporation. She never visits the same branch on consecutive days. If she visits branch  $A$  one day, she visits branch  $B$  the next day. If she visits either branch  $B$  or  $C$  that day, then the next day she is twice as likely to visit branch  $A$  as to visit branch  $B$  or  $C$ . Explain why the resulting transition matrix is regular. Which branch does she visit the most often in the long run?

9.3.3. A study has determined that, on average, a man's occupation depends on that of his father. If the father is a farmer, there is a 30% chance that the son will be a blue collar laborer, a 30% chance he will be a white collar professional, and a 40% chance he will also be a farmer. If the father is a laborer, there is a 30% chance that the son will also be one, a 60% chance he will be a professional, and a 10% chance he will be a farmer. If the father is a professional, there is a 70% chance that the son will also be one, a 25% chance he will be a laborer, and a 5% chance he will be a farmer. (a) What is the probability that the grandson of a farmer will also be a farmer? (b) In the long run, what proportion of the male population will be farmers?

9.3.4. The population of an island is divided into city and country residents. Each year, 5% of the residents of the city move to the country and 15% of the residents of the country move to the city. In 2003, 35,000 people live in the city and 25,000 in the country. Assuming no growth in the population, how many people will live in the city and how many will live in the country between the years 2004 and 2008? What is the eventual population distribution of the island?

9.3.5. A certain plant species has either red, pink, or white flowers, depending on its genotype. If you cross a pink plant with any other plant, the probability distribution of the offspring

is prescribed by the transition matrix  $T = \begin{pmatrix} .5 & .25 & 0 \\ .5 & .5 & .5 \\ 0 & .25 & .5 \end{pmatrix}$ . On average, if you continue

crossing with only pink plants, what percentage of the three types of flowers would you expect to see in your garden?

9.3.6. A genetic model describing inbreeding, in which mating takes place only between individuals of the same genotype, is given by the Markov process  $\mathbf{u}^{(n+1)} = T\mathbf{u}^{(n)}$ ,

where  $T = \begin{pmatrix} 1 & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 1 \end{pmatrix}$  is the transition matrix and  $\mathbf{u}^{(n)} = \begin{pmatrix} p_n \\ q_n \\ r_n \end{pmatrix}$ , whose entries are,

respectively, the proportion of populations of genotype AA, Aa, aa in the  $n^{\text{th}}$  generation. Find the solution to this Markov process and analyze your result.

9.3.7. A student has the habit that if she doesn't study one night, she is 70% certain of studying the next night. Furthermore, the probability that she studies two nights in a row is 50%. How often does she study in the long run?

9.3.8. A traveling salesman visits the three cities of Atlanta, Boston, and Chicago. The matrix

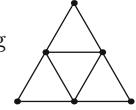
$$\begin{pmatrix} 0 & .5 & .5 \\ 1 & 0 & .5 \\ 0 & .5 & 0 \end{pmatrix}$$

describes the transition probabilities of his trips. Describe his travels in

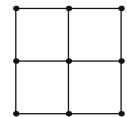
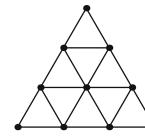
words, and calculate how often he visits each city on average.

9.3.9. Explain why the irregular Markov process with transition matrix  $T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$  does not reach a steady state. Use a population model, as in Exercise 9.3.4, to interpret what is going on.

9.3.10. A bug crawls along the edges of the pictured triangular lattice with six vertices. Upon arriving at a vertex, there is an equal probability of its choosing any edge to leave the vertex. Set up the Markov chain described by the bug's motion, and determine how often, on average, it visits each vertex.



9.3.11. Answer Exercise 9.3.10 for the larger triangular lattice.



9.3.12. Suppose the bug of Exercise 9.3.10 crawls along the edges of the pictured square lattice. What can you say about its behavior?

◊ 9.3.13. Let  $T$  be a regular transition matrix with probability eigenvector  $\mathbf{v}$ .

- (a) Prove that  $\lim_{k \rightarrow \infty} T^k = P = (\mathbf{v} \ \mathbf{v} \ \dots \ \mathbf{v})$  is a matrix with every column equal to  $\mathbf{v}$ .  
 (b) Explain why  $(\mathbf{v} \ \mathbf{v} \ \dots \ \mathbf{v})\mathbf{v} = \mathbf{v}$ . (c) Prove directly that  $P$  is idempotent:  $P^2 = P$ .

9.3.14. Find  $\lim_{k \rightarrow \infty} T^k$  when  $T = \begin{pmatrix} .8 & .1 & .1 \\ .1 & .8 & .1 \\ .1 & .1 & .8 \end{pmatrix}$ .

9.3.15. Prove that, for all  $0 \leq p, q \leq 1$  with  $p + q > 0$ , the probability eigenvector of the transition matrix  $T = \begin{pmatrix} 1-p & q \\ p & 1-q \end{pmatrix}$  is  $\mathbf{v} = \left( \frac{q}{p+q}, \frac{p}{p+q} \right)^T$ .

9.3.16. Describe the final state of a Markov chain with symmetric transition matrix  $T = T^T$ .

9.3.17. *True or false:* If  $T$  and  $T^T$  are both transition matrices, then  $T = T^T$ .

9.3.18. *True or false:* If  $T$  is a transition matrix, so is  $T^{-1}$ .

9.3.19. A transition matrix is called *doubly stochastic* if both its row and column sums are equal to 1. What is the limiting probability state of a Markov chain with doubly stochastic transition matrix?

9.3.20. *True or false:* The set of all probability vectors forms a subspace of  $\mathbb{R}^n$ .

9.3.21. *Multiple choice:* Every probability vector in  $\mathbb{R}^n$  lies on the unit sphere for the  
 (a) 1 norm, (b) 2 norm, (c)  $\infty$  norm, (d) all of the above, (e) none of the above.

9.3.22. *True or false:* Every probability eigenvector of a regular transition matrix has eigenvalue equal to 1.

9.3.23. Write down an example of (a) an irregular transition matrix; (b) a regular transition matrix that has one or more zero entries.

◊ 9.3.24. Let  $T$  be a transition matrix. Prove that if  $\mathbf{u}$  is a probability vector, then so is  $\mathbf{v} = T\mathbf{u}$ .

◊ 9.3.25. (a) Prove that if  $T$  and  $S$  are transition matrices, then so is their product  $TS$ .

- (b) Prove that if  $T$  is a transition matrix, then so is  $T^k$  for all  $k \geq 0$ .

## 9.4 Iterative Solution of Linear Algebraic Systems

In this section, we return to the most basic problem in linear algebra: solving the linear algebraic system

$$A\mathbf{u} = \mathbf{b}, \quad (9.41)$$

consisting of  $n$  equations in  $n$  unknowns. We assume that the  $n \times n$  coefficient matrix  $A$  is nonsingular, and so the solution  $\mathbf{u} = A^{-1}\mathbf{b}$  is unique. For simplicity, we shall only consider real systems here.

We will introduce several popular iterative methods that can be used to approximate the solution for certain classes of coefficient matrices. The resulting algorithms will provide an attractive alternative to Gaussian Elimination, particularly when one is dealing with the large, sparse systems that arise in the numerical solution to differential equations. One major advantage of an iterative technique is that, in favorable situations, it produces progressively more and more accurate approximations to the solution, and hence, by prolonging the iterations, can, at least in principle, compute the solution to any desired order of accuracy. Moreover, even performing just a few iterations may produce a reasonable approximation to the true solution — in stark contrast to Gaussian Elimination, where one must continue the process through to the bitter end before any useful information can be extracted. A partially completed Gaussian Elimination is of scant use! A significant weakness is that iterative methods are not universally applicable, and their design relies upon the detailed structure of the coefficient matrix.

We shall be attempting to solve the linear system (9.41) by replacing it with an iterative system of the form

$$\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}, \quad \mathbf{u}^{(0)} = \mathbf{u}_0, \quad (9.42)$$

in which  $T$  is an  $n \times n$  matrix and  $\mathbf{c} \in \mathbb{R}^n$ . This represents a slight generalization of our earlier iterative system (9.1), in that the right-hand side is now an affine function of  $\mathbf{u}^{(k)}$ . Suppose that the solutions to the affine iterative system converge:  $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^*$  as  $k \rightarrow \infty$ . Then, by taking the limit of both sides of (9.42), we discover that the limit point  $\mathbf{u}^*$  solves the *fixed-point equation*

$$\mathbf{u}^* = T\mathbf{u}^* + \mathbf{c}. \quad (9.43)$$

Thus, we need to design our iterative system so that

- (a) the solution to the fixed-point system  $\mathbf{u} = T\mathbf{u} + \mathbf{c}$  coincides with the solution to the original system  $A\mathbf{u} = \mathbf{b}$ , and
- (b) the iterates defined by (9.42) are known to converge to the fixed point. The more rapid the convergence, the better.

Before exploring these issues in depth, let us look at a simple example.

**Example 9.33.** Consider the linear system

$$3x + y - z = 3, \quad x - 4y + 2z = -1, \quad -2x - y + 5z = 2, \quad (9.44)$$

which has the vectorial form  $A\mathbf{u} = \mathbf{b}$ , with

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}.$$

One easy way to convert a linear system into a fixed-point form is to rewrite it as

$$\mathbf{u} = I\mathbf{u} - A\mathbf{u} + A\mathbf{u} = (I - A)\mathbf{u} + \mathbf{b} = T\mathbf{u} + \mathbf{c}, \quad \text{where} \quad T = I - A, \quad \mathbf{c} = \mathbf{b}.$$

$k$	$\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{b}$			$\mathbf{u}^{(k+1)} = \widehat{T}\mathbf{u}^{(k)} + \widehat{\mathbf{c}}$		
0	0	0	0	0	0	0
1	3	-1	2	1	.25	.4
2	0	-13	-1	1.05	.7	.85
3	15	-64	-7	1.05	.9375	.96
4	30	-322	-4	1.0075	.9925	1.0075
5	261	-1633	-244	1.005	1.00562	1.0015
6	870	-7939	-133	.9986	1.002	1.0031
7	6069	-40300	-5665	1.0004	1.0012	.9999
8	22500	-196240	-5500	.9995	1.0000	1.0004
9	145743	-992701	-129238	1.0001	1.0001	.9998
10	571980	-4850773	-184261	.9999	.9999	1.0001
11	3522555	-24457324	-2969767	1.0000	1.0000	1.0000

In the present case,

$$T = I - A = \begin{pmatrix} -2 & -1 & 1 \\ -1 & 5 & -2 \\ 2 & 1 & -4 \end{pmatrix}, \quad \mathbf{c} = \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}.$$

The resulting iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}$  has the explicit form

$$\begin{aligned} x^{(k+1)} &= -2x^{(k)} - y^{(k)} + z^{(k)} + 3, \\ y^{(k+1)} &= -x^{(k)} + 5y^{(k)} - 2z^{(k)} - 1, \\ z^{(k+1)} &= 2x^{(k)} + y^{(k)} - 4z^{(k)} + 2. \end{aligned} \quad (9.45)$$

Another possibility is to solve the first equation in (9.44) for  $x$ , the second for  $y$ , and the third for  $z$ , so that

$$x = -\frac{1}{3}y + \frac{1}{3}z + 1, \quad y = \frac{1}{4}x + \frac{1}{2}z + \frac{1}{4}, \quad z = \frac{2}{5}x + \frac{1}{5}y + \frac{2}{5}.$$

The resulting equations have the form of a fixed-point system

$$\mathbf{u} = \widehat{T}\mathbf{u} + \widehat{\mathbf{c}}, \quad \text{in which} \quad \widehat{T} = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}, \quad \widehat{\mathbf{c}} = \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{2}{5} \end{pmatrix}.$$

The corresponding iterative system  $\mathbf{u}^{(k+1)} = \widehat{T}\mathbf{u}^{(k)} + \widehat{\mathbf{c}}$  is

$$\begin{aligned} x^{(k+1)} &= -\frac{1}{3}y^{(k)} + \frac{1}{3}z^{(k)} + 1, \\ y^{(k+1)} &= \frac{1}{4}x^{(k)} + \frac{1}{2}z^{(k)} + \frac{1}{4}, \\ z^{(k+1)} &= \frac{2}{5}x^{(k)} + \frac{1}{5}y^{(k)} + \frac{2}{5}. \end{aligned} \quad (9.46)$$

Do the resulting iterative methods converge to the solution  $x = y = z = 1$ , i.e., to  $\mathbf{u}^* = (1, 1, 1)^T$ ? The results, starting with initial guess  $\mathbf{u}^{(0)} = (0, 0, 0)^T$ , are tabulated in the accompanying table.

For the first method, the answer is clearly no — the iterates become wilder and wilder. Indeed, this occurs no matter how close the initial guess  $\mathbf{u}^{(0)}$  is to the actual solution — unless  $\mathbf{u}^{(0)}$  happens to be exactly equal to  $\mathbf{u}^*$ . In the second case, the iterates do converge to the solution, and it does not take too long, even starting from a poor initial guess, to obtain a reasonably accurate approximation. Of course, in such a simple example, it would be silly to use iteration, when Gaussian Elimination can be done by hand and produces the solution almost immediately. However, we use the small examples for illustrative purposes, in order to prepare us to bring the full power of iterative algorithms to bear on the large linear systems arising in applications.

The convergence of solutions to (9.42) to the fixed point  $\mathbf{u}^*$  is based on the behavior of the *error vectors*

$$\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^*, \quad (9.47)$$

which measure how close the iterates are to the true solution. Let us find out how the successive error vectors are related. We compute

$$\mathbf{e}^{(k+1)} = \mathbf{u}^{(k+1)} - \mathbf{u}^* = (T\mathbf{u}^{(k)} + \mathbf{a}) - (T\mathbf{u}^* + \mathbf{a}) = T(\mathbf{u}^{(k)} - \mathbf{u}^*) = T\mathbf{e}^{(k)},$$

showing that the error vectors satisfy a *linear* iterative system

$$\mathbf{e}^{(k+1)} = T\mathbf{e}^{(k)}, \quad (9.48)$$

with the *same* coefficient matrix  $T$ . Therefore, they are given by the explicit formula

$$\mathbf{e}^{(k)} = T^k \mathbf{e}^{(0)}.$$

Now, the solutions to (9.42) converge to the fixed point,  $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^*$ , if and only if the error vectors converge to zero:  $\mathbf{e}^{(k)} \rightarrow \mathbf{0}$  as  $k \rightarrow \infty$ . Our analysis of linear iterative systems, as summarized in Theorem 9.11, establishes the following basic convergence result.

**Proposition 9.34.** The solutions to the affine iterative system (9.42) will all converge to the solution to the fixed point equation (9.43) if and only if  $T$  is a convergent matrix, or, equivalently, its spectral radius satisfies  $\rho(T) < 1$ .

The spectral radius  $\rho(T)$  of the coefficient matrix will govern the speed of convergence. Therefore, our goal is to construct an iterative system whose coefficient matrix has as small a spectral radius as possible. At the very least, the spectral radius must be less than 1. For the two iterative systems presented in Example 9.33, the spectral radii of the coefficient matrices are found to be

$$\rho(T) \simeq 4.9675, \quad \rho(\widehat{T}) = .5.$$

Therefore,  $T$  is not a convergent matrix, which explains the wild behavior of its iterates, whereas  $\widehat{T}$  is convergent, and one expects the error to decrease by a factor of roughly  $\frac{1}{2}$  at each step, which is what is observed in practice.

## The Jacobi Method

The first general iterative method for solving linear systems is based on the same simple idea used in our illustrative Example 9.33. Namely, we solve the  $i^{\text{th}}$  equation in the system  $A\mathbf{u} = \mathbf{b}$ , which is

$$\sum_{j=1}^n a_{ij} u_j = b_i,$$

for the  $i^{\text{th}}$  variable  $u_i$ . To do this, we need to assume that all the diagonal entries of  $A$  are nonzero:  $a_{ii} \neq 0$ . The result is

$$u_i = -\frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} u_j + \frac{b_i}{a_{ii}} = \sum_{j=1}^n t_{ij} u_j + c_i, \quad (9.49)$$

where

$$t_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}}, & i \neq j, \\ 0, & i = j, \end{cases} \quad \text{and} \quad c_i = \frac{b_i}{a_{ii}}. \quad (9.50)$$

The result has the form of a fixed-point system  $\mathbf{u} = T\mathbf{u} + \mathbf{c}$ , and forms the basis of the *Jacobi Method*

$$\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}, \quad \mathbf{u}^{(0)} = \mathbf{u}_0, \quad (9.51)$$

named after the influential nineteenth-century German analyst Carl Jacobi. The explicit form of the Jacobi iterative algorithm is

$$u_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} u_j^{(k)} + \frac{b_i}{a_{ii}}. \quad (9.52)$$

It is instructive to rederive the Jacobi Method in matrix form. We begin by decomposing the coefficient matrix

$$A = L + D + U \quad (9.53)$$

into the sum of a strictly lower triangular matrix  $L$ , meaning all its diagonal entries are 0, a diagonal matrix  $D$ , and a strictly upper triangular matrix  $U$ , each of which is uniquely specified; see Exercise 1.3.11. For example, when

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \quad (9.54)$$

the decomposition (9.53) yields

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -2 & -1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

**Warning.** The  $L, D, U$  in the elementary additive decomposition (9.53) have nothing to do with the  $L, D, U$  appearing in factorizations arising from Gaussian Elimination. The latter play no role in the iterative solution methods considered in this section.

We then rewrite the system

$$A\mathbf{u} = (L + D + U)\mathbf{u} = \mathbf{b} \quad \text{in the alternative form} \quad D\mathbf{u} = -(L + U)\mathbf{u} + \mathbf{b}.$$

The Jacobi fixed point system (9.49) amounts to solving the latter for

$$\mathbf{u} = T\mathbf{u} + \mathbf{c}, \quad \text{where} \quad T = -D^{-1}(L + U), \quad \mathbf{c} = D^{-1}\mathbf{b}. \quad (9.55)$$

For the example (9.54), we recover the Jacobi iteration matrix

$$T = -D^{-1}(L + U) = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}.$$

Deciding in advance whether the Jacobi Method will converge is not easy. However, it can be shown that Jacobi iteration *is* guaranteed to converge when the original coefficient matrix has large diagonal entries, in accordance with Definition 8.18.

**Theorem 9.35.** If the coefficient matrix  $A$  is strictly diagonally dominant, then the associated Jacobi iteration converges.

*Proof:* We shall prove that  $\|T\|_\infty < 1$ , and so Proposition 9.22 implies that  $T$  is a convergent matrix. The absolute row sums of the Jacobi matrix  $T = -D^{-1}(L+U)$  are, according to (9.50),

$$s_i = \sum_{j=1}^n |t_{ij}| = \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}| < 1, \quad (9.56)$$

because  $A$  is strictly diagonally dominant, and hence satisfies (8.28). This implies that  $\|T\|_\infty = \max\{s_1, \dots, s_n\} < 1$ , and the result follows. *Q.E.D.*

**Example 9.36.** Consider the linear system

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

The Jacobi Method solves the respective equations for  $x, y, z, w, v$ , leading to the iterative equations

$$\begin{aligned} x^{(k+1)} &= -\frac{1}{4}y^{(k)} - \frac{1}{4}w^{(k)} + \frac{1}{4}, \\ y^{(k+1)} &= -\frac{1}{4}x^{(k)} - \frac{1}{4}z^{(k)} - \frac{1}{4}v^{(k)} + \frac{1}{2}, \\ z^{(k+1)} &= -\frac{1}{4}y^{(k)} - \frac{1}{4}w^{(k)} - \frac{1}{4}, \\ w^{(k+1)} &= -\frac{1}{4}x^{(k)} - \frac{1}{4}z^{(k)} - \frac{1}{4}v^{(k)} + \frac{1}{2}, \\ v^{(k+1)} &= -\frac{1}{4}y^{(k)} - \frac{1}{4}w^{(k)} + \frac{1}{4}. \end{aligned}$$

The coefficient matrix of the original system,

$$A = \begin{pmatrix} 4 & 1 & 0 & 1 & 0 \\ 1 & 4 & 1 & 0 & 1 \\ 0 & 1 & 4 & 1 & 0 \\ 1 & 0 & 1 & 4 & 1 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix},$$

is strictly diagonally dominant, and so we are guaranteed that the Jacobi iterations will eventually converge to the solution. Indeed, the Jacobi scheme takes the iterative form (9.55), with

$$T = \begin{pmatrix} 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ -\frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix}.$$

Note that  $\|T\|_\infty = \frac{3}{4} < 1$ , validating convergence. Thus, to obtain, say, four decimal place accuracy in the solution, we estimate that it will take fewer than  $\log(.5 \times 10^{-4}) / \log .75 \simeq 34$  iterates, assuming a moderate initial error. But the matrix norm always underestimates the true rate of convergence, as prescribed by the spectral radius  $\rho(T) = .6124$ , which would imply about  $\log(.5 \times 10^{-4}) / \log .6124 \simeq 20$  iterations to obtain the desired accuracy. Indeed, starting with the initial guess  $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$ , the Jacobi iterates converge to the exact solution

$$x = -.1, \quad y = .7, \quad z = -.6, \quad w = .7, \quad v = -.1,$$

to within four decimal places in exactly 20 iterations.

## Exercises

9.4.1.(a) Find the spectral radius of the matrix  $T = \begin{pmatrix} 1 & 1 \\ -1 & -\frac{7}{6} \end{pmatrix}$ . (b) Predict the long term behavior of the iterative system  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{b}$ , where  $\mathbf{b} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ , in as much detail as you can.

9.4.2. Answer Exercise 9.4.1 when (a)  $T = \begin{pmatrix} 1 & -\frac{1}{2} \\ -1 & \frac{3}{2} \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ ;  
 (b)  $T = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \\ 1 & 1 & \frac{1}{4} \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}$ ; (c)  $T = \begin{pmatrix} -0.05 & .15 & .15 \\ .35 & .15 & -.35 \\ -.2 & -.2 & .3 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} -1.5 \\ 1.6 \\ 1.7 \end{pmatrix}$ .

9.4.3. Which of the following systems have a strictly diagonally dominant coefficient matrix?

$$(a) \begin{array}{l} 5x - y = 1, \\ -x + 3y = -1; \end{array} \quad (b) \begin{array}{l} \frac{1}{2}x + \frac{1}{3}y = 1, \\ \frac{1}{5}x + \frac{1}{4}y = 6; \end{array} \quad (c) \begin{array}{l} -5x + y = 3, \\ -3x + 2y = -2; \end{array} \quad (d) \begin{array}{l} -2x + y + z = 1, \\ -x + 2y - z = -2, \\ x - y + 3z = 1; \end{array}$$

$$-x + \frac{1}{2}y + \frac{1}{3}z = 1, \quad x - 2y + z = 1, \quad -4x + 2y + z = 2,$$

$$(e) \begin{array}{l} \frac{1}{3}x + 2y + \frac{3}{4}z = -3, \\ \frac{2}{3}x + \frac{1}{4}y - \frac{3}{2}z = 2; \end{array} \quad (f) \begin{array}{l} -x + 2y + z = -1, \\ x + 3y - 2z = 3; \end{array} \quad (g) \begin{array}{l} -x + 3y + z = -1, \\ x + 4y - 6z = 3. \end{array}$$

♠ 9.4.4. For the strictly diagonally dominant systems in Exercise 9.4.3, starting with the initial guess  $x = y = z = 0$ , compute the solution to 2 decimal places using the Jacobi Method. Check your answer by solving the system directly by Gaussian Elimination.

♠ 9.4.5. (a) Do any of the non-strictly diagonally dominant systems in Exercise 9.4.3 lead to convergent Jacobi algorithms? Hint: Check the spectral radius of the Jacobi matrix.  
 (b) For the convergent systems in Exercise 9.4.3, starting with the initial guess  $x = y = z = 0$ , compute the solution to 2 decimal places by using the Jacobi Method, and check your answer by solving the system directly by Gaussian Elimination.

9.4.6. The following linear systems have positive definite coefficient matrices. Use the Jacobi Method starting with  $\mathbf{u}^{(0)} = \mathbf{0}$  to find the solution to 4 decimal place accuracy.

$$(a) \begin{pmatrix} 3 & -1 \\ -1 & 5 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{u} = \begin{pmatrix} -3 \\ 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 6 & -1 & -3 \\ -1 & 7 & 4 \\ -3 & 4 & 9 \end{pmatrix} \mathbf{u} = \begin{pmatrix} -1 \\ -2 \\ 7 \end{pmatrix},$$

$$(d) \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 5 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 1 \\ -5 \\ 0 \end{pmatrix}, \quad (e) \begin{pmatrix} 5 & 1 & 1 & 1 \\ 1 & 5 & 1 & 1 \\ 1 & 1 & 5 & 1 \\ 1 & 1 & 1 & 5 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 4 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad (f) \begin{pmatrix} 3 & 1 & 0 & -1 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ -1 & 0 & 1 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}.$$

- ♣ 9.4.7. Let  $A$  be the  $n \times n$  tridiagonal matrix with all its diagonal entries equal to  $c$  and all 1's on the sub- and super-diagonals. (a) For which values of  $c$  is  $A$  strictly diagonally dominant? (b) For which values of  $c$  does the Jacobi iteration for  $A\mathbf{u} = \mathbf{b}$  converge to the solution? What is the rate of convergence? *Hint:* Use Exercise 8.2.48. (c) Set  $c = 2$  and use the Jacobi Method to solve the linear systems  $K\mathbf{u} = \mathbf{e}_1$ , for  $n = 5, 10$ , and  $20$ . Starting with an initial guess of  $\mathbf{0}$ , how many Jacobi iterations does it take to obtain 3 decimal place accuracy? Does the convergence rate agree with what you computed in part (c)?
- 9.4.8. Prove that  $\mathbf{0} \neq \mathbf{u} \in \ker A$  if and only if  $\mathbf{u}$  is an eigenvector of the Jacobi iteration matrix with eigenvalue 1. What does this imply about convergence?
- ◇ 9.4.9. Prove that if  $A$  is a nonsingular coefficient matrix, then one can always arrange that all its diagonal entries are nonzero by suitably permuting its rows.
- 9.4.10. Consider the iterative system (9.42) with spectral radius  $\rho(T) < 1$ . Explain why it takes roughly  $-1/\log_{10} \rho(T)$  iterations to produce one further decimal digit of accuracy in the solution.
- 9.4.11. *True or false:* If a system  $A\mathbf{u} = \mathbf{b}$  has a strictly diagonally dominant coefficient matrix  $A$ , then the equivalent system obtained by applying an elementary row operation to  $A$  also has a strictly diagonally dominant coefficient matrix.

## The Gauss–Seidel Method

The Gauss–Seidel Method relies on a slightly refined implementation of the Jacobi process. To understand how it works, it will help to write out the Jacobi iteration algorithm (9.51) in full detail:

$$\begin{aligned}
 u_1^{(k+1)} &= t_{12}u_2^{(k)} + t_{13}u_3^{(k)} + \cdots + t_{1,n-1}u_{n-1}^{(k)} + t_{1n}u_n^{(k)} + c_1, \\
 u_2^{(k+1)} &= t_{21}u_1^{(k)} + t_{23}u_3^{(k)} + \cdots + t_{2,n-1}u_{n-1}^{(k)} + t_{2n}u_n^{(k)} + c_2, \\
 u_3^{(k+1)} &= t_{31}u_1^{(k)} + t_{32}u_2^{(k)} + \cdots + t_{3,n-1}u_{n-1}^{(k)} + t_{3n}u_n^{(k)} + c_3, \\
 &\vdots \quad \vdots \quad \vdots \quad \ddots \quad \ddots \quad \vdots \\
 u_n^{(k+1)} &= t_{n1}u_1^{(k)} + t_{n2}u_2^{(k)} + t_{n3}u_3^{(k)} + \cdots + t_{n,n-1}u_{n-1}^{(k)} + c_n,
 \end{aligned} \tag{9.57}$$

where we are explicitly noting the fact that all the diagonal entries of the coefficient matrix  $T$  vanish. Observe that we are using the entries of the current iterate  $\mathbf{u}^{(k)}$  to compute *all* of the updated values of  $\mathbf{u}^{(k+1)}$ . Presumably, if the iterates  $\mathbf{u}^{(k)}$  are converging to the solution  $\mathbf{u}^*$ , then their individual entries are also converging, and so each  $u_j^{(k+1)}$  should be a better approximation to  $u_j^*$  than  $u_j^{(k)}$  is. Therefore, if we begin the  $k^{\text{th}}$  Jacobi iteration by computing  $u_1^{(k+1)}$  using the first equation, then we are tempted to use this new and improved value to replace  $u_1^{(k)}$  in each of the subsequent equations. In particular, we employ the modified equation

$$u_2^{(k+1)} = t_{21}u_1^{(k+1)} + t_{23}u_3^{(k)} + \cdots + t_{1n}u_n^{(k)} + c_2$$

to update the second component of our iterate. This more accurate value should then be used to update  $u_3^{(k+1)}$ , and so on.

The upshot of these considerations is the *Gauss–Seidel Method*

$$u_i^{(k+1)} = t_{i1}u_1^{(k+1)} + \cdots + t_{i,i-1}u_{i-1}^{(k+1)} + t_{i,i+1}u_{i+1}^{(k)} + \cdots + t_{in}u_n^{(k)} + c_i, \quad i = 1, \dots, n, \tag{9.58}$$

named after Gauss (as usual!) and the German astronomer/mathematician Philipp von Seidel. At the  $k^{\text{th}}$  stage of the iteration, we use (9.58) to compute the revised entries  $u_1^{(k+1)}, u_2^{(k+1)}, \dots, u_n^{(k+1)}$  in their numerical order. Once an entry has been updated, the new value is immediately used in all subsequent computations.

**Example 9.37.** For the linear system

$$3x + y - z = 3, \quad x - 4y + 2z = -1, \quad -2x - y + 5z = 2,$$

the Jacobi iteration method was given in (9.46). To construct the corresponding Gauss–Seidel algorithm we use updated values of  $x, y$ , and  $z$  as they become available. Explicitly,

$$\begin{aligned} x^{(k+1)} &= -\frac{1}{3}y^{(k)} + \frac{1}{3}z^{(k)} + 1, \\ y^{(k+1)} &= \frac{1}{4}x^{(k+1)} + \frac{1}{2}z^{(k)} + \frac{1}{4}, \\ z^{(k+1)} &= \frac{2}{5}x^{(k+1)} + \frac{1}{5}y^{(k+1)} + \frac{2}{5}. \end{aligned} \quad (9.59)$$

Starting with  $\mathbf{u}^{(0)} = \mathbf{0}$ , the resulting iterates are

$$\begin{aligned} \mathbf{u}^{(1)} &= \begin{pmatrix} 1.0000 \\ .5000 \\ .9000 \end{pmatrix}, & \mathbf{u}^{(2)} &= \begin{pmatrix} 1.1333 \\ .9833 \\ 1.0500 \end{pmatrix}, & \mathbf{u}^{(3)} &= \begin{pmatrix} 1.0222 \\ 1.0306 \\ 1.0150 \end{pmatrix}, & \mathbf{u}^{(4)} &= \begin{pmatrix} .9948 \\ 1.0062 \\ .9992 \end{pmatrix}, \\ \mathbf{u}^{(5)} &= \begin{pmatrix} .9977 \\ .9990 \\ .9989 \end{pmatrix}, & \mathbf{u}^{(6)} &= \begin{pmatrix} 1.0000 \\ .9994 \\ .9999 \end{pmatrix}, & \mathbf{u}^{(7)} &= \begin{pmatrix} 1.0001 \\ 1.0000 \\ 1.0001 \end{pmatrix}, & \mathbf{u}^{(8)} &= \begin{pmatrix} 1.0000 \\ 1.0000 \\ 1.0000 \end{pmatrix}, \end{aligned}$$

and have converged to the solution, to 4 decimal place accuracy, after only 8 iterations — as opposed to the 11 iterations required by the Jacobi Method.

Gauss–Seidel iteration is particularly suited to implementation on a serial computer, since one can immediately replace each component  $u_i^{(k)}$  by its updated value  $u_i^{(k+1)}$ , thereby also saving on storage in the computer’s memory. In contrast, the Jacobi Method requires us to retain all the old values  $\mathbf{u}^{(k)}$  until the new approximation  $\mathbf{u}^{(k+1)}$  has been computed. Moreover, Gauss–Seidel typically (although not always) converges faster than Jacobi, making it the iterative algorithm of choice for serial processors. On the other hand, with the advent of parallel processing machines, variants of the parallelizable Jacobi scheme have been making a comeback.

What is Gauss–Seidel really up to? Let us rewrite the basic iterative equation (9.58) by multiplying by  $a_{ii}$  and moving the terms involving  $\mathbf{u}^{(k+1)}$  to the left-hand side. In view of the formula (9.50) for the entries of  $T$ , the resulting equation is

$$a_{i1}u_1^{(k+1)} + \dots + a_{i,i-1}u_{i-1}^{(k+1)} + a_{ii}u_i^{(k+1)} = -a_{i,i+1}u_{i+1}^{(k)} - \dots - a_{in}u_n^{(k)} + b_i.$$

In matrix form, taking (9.53) into account, this reads

$$(L + D)\mathbf{u}^{(k+1)} = -U\mathbf{u}^{(k)} + \mathbf{b}, \quad (9.60)$$

and so can be viewed as a linear system of equations for  $\mathbf{u}^{(k+1)}$  with lower triangular coefficient matrix  $L + D$ . Note that the fixed point of (9.60), namely the solution to

$$(L + D)\mathbf{u} = -U\mathbf{u} + \mathbf{b},$$

coincides with the solution to the original system

$$A\mathbf{u} = (L + D + U)\mathbf{u} = \mathbf{b}.$$

In other words, the Gauss–Seidel procedure is merely implementing Forward Substitution to solve the lower triangular system (9.60) for the next iterate:

$$\mathbf{u}^{(k+1)} = -(L + D)^{-1}U\mathbf{u}^{(k)} + (L + D)^{-1}\mathbf{b}.$$

The latter is in our more usual iterative form

$$\mathbf{u}^{(k+1)} = \tilde{T}\mathbf{u}^{(k)} + \tilde{\mathbf{c}}, \quad \text{where} \quad \tilde{T} = -(L + D)^{-1}U, \quad \tilde{\mathbf{c}} = (L + D)^{-1}\mathbf{b}. \quad (9.61)$$

Consequently, the convergence of the Gauss–Seidel iterates is governed by the spectral radius of their coefficient matrix  $\tilde{T}$ .

Returning to Example 9.37, we have

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \quad L + D = \begin{pmatrix} 3 & 0 & 0 \\ 1 & -4 & 0 \\ -2 & -1 & 5 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, the Gauss–Seidel matrix is

$$\tilde{T} = -(L + D)^{-1}U = \begin{pmatrix} 0 & -.3333 & .3333 \\ 0 & -.0833 & .5833 \\ 0 & -.1500 & .2500 \end{pmatrix}.$$

Its eigenvalues are 0 and  $.0833 \pm .2444i$ , and hence its spectral radius is  $\rho(\tilde{T}) \simeq .2582$ . This is roughly the square of the Jacobi spectral radius of .5, which tells us that the Gauss–Seidel iterations will converge about twice as fast to the solution. This can be verified by more extensive computations. Although examples can be constructed in which the Jacobi Method converges faster, in many practical situations Gauss–Seidel tends to converge roughly twice as fast as Jacobi.

Completely general conditions guaranteeing convergence of the Gauss–Seidel Method are also hard to establish. But, like the Jacobi Method, it is guaranteed to converge when the original coefficient matrix is strictly diagonally dominant.

**Theorem 9.38.** If  $A$  is strictly diagonally dominant, then the Gauss–Seidel iteration algorithm for solving  $A\mathbf{u} = \mathbf{b}$  converges.

*Proof:* Let  $\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^*$  denote the  $k^{\text{th}}$  Gauss–Seidel error vector. As in (9.48), the error vectors satisfy the linear iterative system  $\mathbf{e}^{(k+1)} = \tilde{T}\mathbf{e}^{(k)}$ , but a direct estimate of  $\|\tilde{T}\|_\infty$  is not so easy. Instead, let us write out the linear iterative system in components:

$$e_i^{(k+1)} = t_{i1}e_1^{(k+1)} + \dots + t_{i,i-1}e_{i-1}^{(k+1)} + t_{i,i+1}e_{i+1}^{(k)} + \dots + t_{in}e_n^{(k)}. \quad (9.62)$$

Let

$$m^{(k)} = \|\mathbf{e}^{(k)}\|_\infty = \max\{|e_1^{(k)}|, \dots, |e_n^{(k)}|\} \quad (9.63)$$

denote the  $\infty$  norm of the  $k^{\text{th}}$  error vector. To prove convergence,  $\mathbf{e}^{(k)} \rightarrow \mathbf{0}$ , it suffices to

show that  $m^{(k)} \rightarrow 0$  as  $k \rightarrow \infty$ . We claim that diagonal dominance of  $A$  implies that

$$m^{(k+1)} \leq s m^{(k)}, \quad \text{where} \quad s = \|T\|_{\infty} < 1 \quad (9.64)$$

denotes the  $\infty$  matrix norm of the *Jacobi* matrix  $T$  — not the Gauss–Seidel matrix  $\tilde{T}$  — which, by (9.56), is less than 1. We infer that  $m^{(k)} \leq s^k m^{(0)} \rightarrow 0$  as  $k \rightarrow \infty$ , demonstrating the theorem.

To prove (9.64), we use induction on  $i = 1, \dots, n$ . Our induction hypothesis is

$$|e_j^{(k+1)}| \leq s m^{(k)} < m^{(k)} \quad \text{for} \quad j = 1, \dots, i-1.$$

(When  $i = 1$ , there is no assumption.) Moreover, by (9.63),

$$|e_j^{(k)}| \leq m^{(k)} \quad \text{for all} \quad j = 1, \dots, n.$$

We use these two inequalities to estimate  $|e_i^{(k+1)}|$  from (9.62):

$$\begin{aligned} |e_i^{(k+1)}| &\leq |t_{i1}| |e_1^{(k+1)}| + \dots + |t_{i,i-1}| |e_{i-1}^{(k+1)}| + |t_{i,i+1}| |e_{i+1}^{(k)}| + \dots + |t_{in}| |e_n^{(k)}| \\ &\leq (|t_{i1}| + \dots + |t_{in}|) m^{(k)} \leq s m^{(k)}, \end{aligned}$$

which completes the induction step. As a result, the maximum

$$m^{(k+1)} = \max\{|e_1^{(k+1)}|, \dots, |e_n^{(k+1)}|\} \leq s m^{(k)}$$

also satisfies the same bound, and hence (9.64) follows. *Q.E.D.*

**Example 9.39.** For the linear system considered in Example 9.36, the Gauss–Seidel iterations take the form

$$\begin{aligned} x^{(k+1)} &= -\frac{1}{4}y^{(k)} - \frac{1}{4}w^{(k)} + \frac{1}{4}, \\ y^{(k+1)} &= -\frac{1}{4}x^{(k+1)} - \frac{1}{4}z^{(k)} - \frac{1}{4}v^{(k)} + \frac{1}{2}, \\ z^{(k+1)} &= -\frac{1}{4}y^{(k+1)} - \frac{1}{4}w^{(k)} - \frac{1}{4}, \\ w^{(k+1)} &= -\frac{1}{4}x^{(k+1)} - \frac{1}{4}z^{(k+1)} - \frac{1}{4}v^{(k)} + \frac{1}{2}, \\ v^{(k+1)} &= -\frac{1}{4}y^{(k+1)} - \frac{1}{4}w^{(k+1)} + \frac{1}{4}. \end{aligned}$$

Starting with  $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$ , the Gauss–Seidel iterates converge to the solution  $x = -.1$ ,  $y = .7$ ,  $z = -.6$ ,  $w = .7$ ,  $v = -.1$ , to four decimal places in 11 iterations, again roughly twice as fast as the Jacobi Method. Indeed, the convergence rate is governed by the corresponding Gauss–Seidel matrix  $\tilde{T}$ , which is

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -.2500 & 0 & -.2500 & 0 \\ 0 & .0625 & -.2500 & .0625 & -.2500 \\ 0 & -.0156 & .0625 & -.2656 & .0625 \\ 0 & .0664 & -.0156 & .1289 & -.2656 \\ 0 & -.0322 & .0664 & -.0479 & .1289 \end{pmatrix}.$$

Its spectral radius is  $\rho(\tilde{T}) = .3936$ , which is, as in the previous example, approximately the square of the spectral radius of the Jacobi coefficient matrix, which explains the speedup in convergence.

## Exercises

- 9.4.12. Consider the linear system  $A\mathbf{x} = \mathbf{b}$ , where  $A = \begin{pmatrix} 4 & 1 & -2 \\ -1 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} -2 \\ -1 \\ 7 \end{pmatrix}$ .
- (a) First, solve the equation directly by Gaussian Elimination. (b) Write the Jacobi iteration in the form  $\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + \mathbf{c}$ . Find the  $3 \times 3$  matrix  $T$  and the vector  $\mathbf{c}$  explicitly. (c) Using the initial approximation  $\mathbf{x}^{(0)} = \mathbf{0}$ , carry out three iterations of the Jacobi algorithm to compute  $\mathbf{x}^{(1)}$ ,  $\mathbf{x}^{(2)}$  and  $\mathbf{x}^{(3)}$ . How close are you to the exact solution? (d) Write the Gauss–Seidel iteration in the form  $\mathbf{x}^{(k+1)} = \tilde{T}\mathbf{x}^{(k)} + \tilde{\mathbf{c}}$ . Find the  $3 \times 3$  matrix  $\tilde{T}$  and the vector  $\tilde{\mathbf{c}}$  explicitly. (e) Using the initial approximation  $\mathbf{x}^{(0)} = \mathbf{0}$ , carry out three iterations of the Gauss–Seidel algorithm. Which is a better approximation to the solution — Jacobi or Gauss–Seidel? (f) Determine the spectral radius of the Jacobi matrix  $T$ , and use this to prove that the Jacobi Method will converge to the solution of  $A\mathbf{x} = \mathbf{b}$  for any choice of the initial approximation  $\mathbf{x}^{(0)}$ . (g) Determine the spectral radius of the Gauss–Seidel matrix  $\tilde{T}$ . Which method converges faster? (h) For the faster method, how many iterations would you expect to need to obtain 5 decimal place accuracy? (i) Test your prediction by computing the solution to the desired accuracy.
- 9.4.13. For the strictly diagonally dominant systems in Exercise 9.4.3, starting with the initial guess  $x = y = z = 0$ , compute the solution to 3 decimal places using the Gauss–Seidel Method. Check your answer by solving the system directly by Gaussian Elimination.
- 9.4.14. Which of the systems in Exercise 9.4.3 lead to convergent Gauss–Seidel algorithms? In each case, which converges faster, Jacobi or Gauss–Seidel?
- 9.4.15. (a) Solve the positive definite linear systems in Exercise 9.4.6 using the Gauss–Seidel Method to achieve 4 decimal place accuracy.  
 (b) Compare the convergence rate with that of the Jacobi Method.
- 9.4.16. Let  $A = \begin{pmatrix} c & 1 & 0 & 0 \\ 1 & c & 1 & 0 \\ 0 & 1 & c & 1 \\ 0 & 0 & 1 & c \end{pmatrix}$ . (a) For what values of  $c$  is  $A$  strictly diagonally dominant?  
 (b) Use a computer to find the smallest positive value of  $c > 0$  for which Jacobi iteration converges. (c) Find the smallest positive value of  $c > 0$  for which Gauss–Seidel iteration converges. Is your answer the same? (d) When they both converge, which converges faster — Jacobi or Gauss–Seidel? How much faster? Does your answer depend upon the value of  $c$ ?
- 9.4.17. Consider the linear system  

$$2.4x - .8y + .8z = 1, \quad -.6x + 3.6y - .6z = 0, \quad 15x + 14.4y - 3.6z = 0.$$
 Show, by direct computation, that Jacobi iteration converges to the solution, but Gauss–Seidel does not.
- 9.4.18. Discuss convergence of Gauss–Seidel iteration for the system  

$$\begin{aligned} 5x + 7y + 6z + 5w &= 23, & 6x + 8y + 10z + 9w &= 33, \\ 7x + 10y + 8z + 7w &= 32, & 5x + 7y + 9z + 10w &= 31. \end{aligned}$$
- 9.4.19. Let  $A = \begin{pmatrix} 2 & 4 & -4 \\ 3 & 3 & 3 \\ 2 & 2 & 1 \end{pmatrix}$ . Find the spectral radius of the Jacobi and Gauss–Seidel iteration matrices, and discuss their convergence.
- 9.4.20. Consider the linear system  $H_5\mathbf{u} = \mathbf{e}_1$ , where  $H_5$  is the  $5 \times 5$  Hilbert matrix. Does the Jacobi Method converge to the solution? If so, how fast? What about Gauss–Seidel?

- ◊ 9.4.21. How many arithmetic operations are needed to perform  $k$  steps of the Jacobi iteration? What about Gauss–Seidel? Under what conditions is Jacobi or Gauss–Seidel more efficient than Gaussian Elimination?

- ♣ 9.4.22. Consider the linear system  $A\mathbf{x} = \mathbf{e}_1$  based on the  $10 \times 10$  pentadiagonal matrix

$$A = \begin{pmatrix} z & -1 & 1 & 0 & & & & & & \\ -1 & z & -1 & 1 & 0 & & & & & \\ 1 & -1 & z & -1 & 1 & 0 & & & & \\ 0 & 1 & -1 & z & -1 & 1 & \ddots & & & \\ & 0 & 1 & -1 & z & -1 & \ddots & & & \\ & & 0 & 1 & -1 & z & \ddots & & & \\ & & & \ddots & \ddots & \ddots & \ddots & & & \end{pmatrix}.$$

- (a) For what values of  $z$  are the Jacobi and Gauss–Seidel Methods guaranteed to converge?  
 (b) Set  $z = 4$ . How many iterations are required to approximate the solution to 3 decimal places?  
 (c) How small can  $|z|$  be before the methods diverge?

- ♣ 9.4.23. The *naïve iterative method* for solving  $A\mathbf{u} = \mathbf{b}$  is to rewrite it in fixed point form  $\mathbf{u} = T\mathbf{u} + \mathbf{c}$ , where  $T = I - A$  and  $\mathbf{c} = \mathbf{b}$ . (a) What conditions on the eigenvalues of  $A$  ensure convergence of the naïve method? (b) Use the Gershgorin Theorem 8.16 to prove

that the naïve method converges to the solution to  $\begin{pmatrix} .8 & -.1 & -.1 \\ .2 & 1.5 & -.1 \\ .2 & -.1 & 1.0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ .  
 (c) Check part (b) by implementing the method.

## Successive Over-Relaxation

As we know, the smaller the spectral radius (or matrix norm) of the coefficient matrix, the faster the convergence of the iterative algorithm. One of the goals of researchers in numerical linear algebra is to design new methods for accelerating the convergence. In his 1950 thesis, the American mathematician David Young discovered a simple modification of the Jacobi and Gauss–Seidel Methods that can, in favorable situations, lead to a dramatic speedup in the rate of convergence. The method, known as *Successive Over-Relaxation*, and often abbreviated SOR, has become the iterative method of choice in a range of modern applications, [21, 86]. In this subsection, we provide a brief overview.

In practice, finding the optimal iterative algorithm to solve a given linear system is as hard as solving the system itself. Therefore, numerical analysts have relied on a few tried and true techniques for designing iterative schemes that can be used in the more common applications. Consider a linear algebraic system  $A\mathbf{u} = \mathbf{b}$ . Every decomposition of the coefficient matrix into the difference of two matrices,

$$A = M - N, \tag{9.65}$$

leads to an equivalent system of the form

$$M\mathbf{u} = N\mathbf{u} + \mathbf{b}. \tag{9.66}$$

Provided that  $M$  is nonsingular, we can rewrite the preceding system in fixed point form:

$$\mathbf{u} = M^{-1}N\mathbf{u} + M^{-1}\mathbf{b} = T\mathbf{u} + \mathbf{c}, \quad \text{where} \quad T = M^{-1}N, \quad \mathbf{c} = M^{-1}\mathbf{b}.$$

Now, we are free to choose any such  $M$ , which then specifies  $N = A - M$  uniquely. However, for the resulting iterative method  $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}$  to be practical we must arrange that

- (a)  $T = M^{-1}N$  is a convergent matrix, and  
 (b)  $M$  can be easily inverted.

The second requirement ensures that the iterative equations

$$M \mathbf{u}^{(k+1)} = N \mathbf{u}^{(k)} + \mathbf{b} \quad (9.67)$$

can be solved for  $\mathbf{u}^{(k+1)}$  with minimal computational effort. Typically, this requires that  $M$  be either a diagonal matrix, in which case the inversion is immediate, or lower or upper triangular, in which case one employs Forward or Back Substitution to solve for  $\mathbf{u}^{(k+1)}$ .

With this in mind, we now introduce the SOR Method. It relies on a slight generalization of the Gauss–Seidel decomposition (9.60) of the matrix into lower triangular and strictly upper triangular parts. The starting point is to write

$$A = L + D + U = [L + \alpha D] - [(\alpha - 1)D - U], \quad (9.68)$$

where  $0 \neq \alpha$  is an adjustable scalar parameter. We decompose the system  $A \mathbf{u} = \mathbf{b}$  as

$$(L + \alpha D)\mathbf{u} = [(\alpha - 1)D - U]\mathbf{u} + \mathbf{b}. \quad (9.69)$$

It turns out to be slightly more convenient to divide (9.69) through by  $\alpha$  and write the resulting iterative system in the form

$$(\omega L + D)\mathbf{u}^{(k+1)} = [(1 - \omega)D - \omega U]\mathbf{u}^{(k)} + \omega \mathbf{b}, \quad (9.70)$$

where  $\omega = 1/\alpha$  is called the *relaxation parameter*. Assuming, as usual, that all diagonal entries of  $A$  are nonzero, the matrix  $\omega L + D$  is an invertible lower triangular matrix, and so we can use Forward Substitution to solve the iterative system (9.70) to recover  $\mathbf{u}^{(k+1)}$ . The explicit formula for its  $i^{\text{th}}$  entry is

$$\begin{aligned} u_i^{(k+1)} &= \omega t_{i1} u_1^{(k+1)} + \dots + \omega t_{i,i-1} u_{i-1}^{(k+1)} + (1 - \omega) u_i^{(k)} \\ &\quad + \omega t_{i,i+1} u_{i+1}^{(k)} + \dots + \omega t_{in} u_n^{(k)} + \omega c_i, \end{aligned} \quad (9.71)$$

where  $t_{ij}$  and  $c_i$  denote the original Jacobi values (9.50). As in the Gauss–Seidel approach, we update the entries  $u_i^{(k+1)}$  in numerical order  $i = 1, \dots, n$ . Thus, to obtain the SOR scheme (9.71), we merely multiply the right-hand side of the Gauss–Seidel system (9.58) by the adjustable relaxation parameter  $\omega$  and append the diagonal term  $(1 - \omega)u_i^{(k)}$ . In particular, if we set  $\omega = 1$ , then the SOR Method reduces to the Gauss–Seidel Method. Choosing  $\omega < 1$  leads to an *under-relaxed* method, while  $\omega > 1$ , known as *over-relaxation*, is the preferred choice in most practical instances.

To analyze the SOR algorithm in detail, we rewrite (9.70) in the fixed point form

$$\mathbf{u}^{(k+1)} = T_\omega \mathbf{u}^{(k)} + \mathbf{c}_\omega, \quad (9.72)$$

where

$$T_\omega = (\omega L + D)^{-1} [(\alpha - 1)D - \omega U], \quad \mathbf{c}_\omega = (\omega L + D)^{-1} \omega \mathbf{b}. \quad (9.73)$$

The rate of convergence is governed by the spectral radius of the matrix  $T_\omega$ . The goal is to choose the relaxation parameter  $\omega$  so as to make the spectral radius of  $T_\omega$  as small as possible. As we will see, a clever choice of  $\omega$  can result in a dramatic speedup in the convergence rate. Let us look at an elementary example.

**Example 9.40.** Consider the matrix  $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ , which we decompose as  $A = L + D + U$ , where

$$L = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

Jacobi iteration is based on the coefficient matrix

$$T = -D^{-1}(L + U) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}.$$

Its spectral radius is  $\rho(T) = .5$ , and hence the Jacobi Method takes, on average, roughly  $-1/\log_{10}.5 \simeq 3.3$  iterations to produce each new decimal place in the solution.

The SOR Method (9.70) takes the explicit form

$$\begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix} \mathbf{u}^{(k+1)} = \begin{pmatrix} 2(1-\omega) & \omega \\ 0 & 2(1-\omega) \end{pmatrix} \mathbf{u}^{(k)} + \omega \mathbf{b},$$

where Gauss–Seidel is the particular case  $\omega = 1$ . The SOR coefficient matrix is

$$T_\omega = \begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2(1-\omega) & \omega \\ 0 & 2(1-\omega) \end{pmatrix} = \begin{pmatrix} 1-\omega & \frac{1}{2}\omega \\ \frac{1}{2}\omega(1-\omega) & \frac{1}{4}(2-\omega)^2 \end{pmatrix}.$$

To compute the eigenvalues of  $T_\omega$ , we form its characteristic equation:

$$0 = \det(T_\omega - \lambda \mathbf{I}) = \lambda^2 - (2 - 2\omega + \frac{1}{4}\omega^2)\lambda + (1-\omega)^2 = (\lambda + \omega - 1)^2 - \frac{1}{4}\lambda\omega^2. \quad (9.74)$$

Our goal is to choose  $\omega$  such that

- (a) both eigenvalues are less than 1 in modulus, so  $|\lambda_1|, |\lambda_2| < 1$ . This is the minimal requirement for convergence of the method.
- (b) the largest eigenvalue (in modulus) is as small as possible. This will give the smallest spectral radius for  $T_\omega$  and hence the fastest convergence rate.

By (8.26), the product of the two eigenvalues is the determinant,

$$\lambda_1 \lambda_2 = \det T_\omega = (1-\omega)^2.$$

If  $\omega \leq 0$  or  $\omega \geq 2$ , then  $\det T_\omega \geq 1$ , and hence at least one of the eigenvalues would have modulus larger than 1. Thus, in order to ensure convergence, we must require  $0 < \omega < 2$ . For Gauss–Seidel, at  $\omega = 1$ , the eigenvalues are  $\lambda_1 = \frac{1}{4}$ ,  $\lambda_2 = 0$ , and the spectral radius is  $\rho(T_1) = .25$ . This is exactly the square of the Jacobi spectral radius, and hence the Gauss–Seidel iterates converge twice as fast; so it takes, on average, only about  $-1/\log_{10}.25 \simeq 1.66$  Gauss–Seidel iterations to produce each new decimal place of accuracy. It can be shown (Exercise 9.4.32) that as  $\omega$  increases above 1, the two eigenvalues move along the real axis towards each other. They coincide when

$$\omega = \omega_* = 8 - 4\sqrt{3} \simeq 1.07, \quad \text{at which point} \quad \lambda_1 = \lambda_2 = \omega_* - 1 = .07 = \rho(T_\omega),$$

which is the convergence rate of the optimal SOR Method. Each iteration produces slightly more than one new decimal place in the solution, which represents a significant improvement over the Gauss–Seidel convergence rate. It takes about twice as many Gauss–Seidel iterations (and four times as many Jacobi iterations) to produce the same accuracy as this optimal SOR Method.

Of course, in such a simple  $2 \times 2$  example, it is not so surprising that we can construct the best value for the relaxation parameter by hand. Young was able to find the optimal value of the relaxation parameter for a broad class of matrices that includes most of those arising in the finite difference and finite element numerical solutions to ordinary and partial differential equations, [61]. For the matrices in Young’s class, the Jacobi eigenvalues

occur in signed pairs. If  $\pm\mu$  are a pair of eigenvalues for the Jacobi Method, then the corresponding eigenvalues of the SOR iteration matrix satisfy the quadratic equation

$$(\lambda + \omega - 1)^2 = \lambda \omega^2 \mu^2. \quad (9.75)$$

If  $\omega = 1$ , so we have standard Gauss–Seidel, then  $\lambda^2 = \lambda \mu^2$ , and so the eigenvalues are  $\lambda = 0$ ,  $\lambda = \mu^2$ . The Gauss–Seidel spectral radius is therefore the square of the Jacobi spectral radius, and so (at least for matrices in the Young class) its iterates converge twice as fast. The quadratic equation (9.75) has the same properties as in the  $2 \times 2$  version (9.74) (which corresponds to the case  $\mu = \frac{1}{2}$ ), and hence the optimal value of  $\omega$  will be the one at which the two roots are equal:

$$\lambda_1 = \lambda_2 = \omega - 1, \quad \text{which occurs when} \quad \omega = \frac{2 - 2\sqrt{1 - \mu^2}}{\mu^2} = \frac{2}{1 + \sqrt{1 - \mu^2}}.$$

Therefore, if  $\rho_J = \max |\mu|$  denotes the spectral radius of the Jacobi Method, then the Gauss–Seidel has spectral radius  $\rho_{GS} = \rho_J^2$ , while the SOR Method with optimal relaxation parameter

$$\omega_* = \frac{2}{1 + \sqrt{1 - \rho_J^2}}, \quad \text{has spectral radius} \quad \rho_* = \omega_* - 1. \quad (9.76)$$

For example, if  $\rho_J = .99$ , which is rather slow convergence (but common for iterative numerical solution schemes for partial differential equations), then  $\rho_{GS} = .9801$ , which is twice as fast, but still quite slow, while SOR with  $\omega_* = 1.7527$  has  $\rho_* = .7527$ , which is dramatically faster<sup>†</sup>. Indeed, since  $\rho_* \simeq (\rho_{GS})^{14} \simeq (\rho_J)^{28}$ , it takes about 14 Gauss–Seidel (and 28 Jacobi) iterations to produce the same accuracy as one SOR step. It is amazing that such a simple idea can have such a dramatic effect.

## Exercises

♡ 9.4.24. Consider the linear system  $A\mathbf{u} = \mathbf{b}$ , where  $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ .

- (a) What is the solution? (b) Discuss the convergence of the Jacobi iteration method.
- (c) Discuss the convergence of the Gauss–Seidel iteration method. (d) Write down the explicit formulas for the SOR Method. (e) What is the optimal value of the relaxation parameter  $\omega$  for this system? How much faster is the convergence as compared to the Jacobi and Gauss–Seidel Methods? (f) Suppose your initial guess is  $\mathbf{u}^{(0)} = \mathbf{0}$ . Give an estimate as to how many steps each iterative method (Jacobi, Gauss–Seidel, SOR) would require in order to approximate the solution to the system to within 5 decimal places.
- (g) Verify your answer by direct computation.

♠ 9.4.25. In Exercise 9.4.18 you were asked to solve a system by Gauss–Seidel. How much faster can you design an SOR scheme to converge? Experiment with several values of the relaxation parameter  $\omega$ , and discuss what you find.

♠ 9.4.26. Investigate the three basic iterative techniques — Jacobi, Gauss–Seidel, SOR — for solving the linear system  $K^* \mathbf{u}^* = \mathbf{f}^*$  for the cubical circuit in Example 6.4.

<sup>†</sup> More precisely, since the SOR matrix is not necessarily diagonalizable, the overall convergence rate is slightly slower than the spectral radius. However, this technical detail does not affect the overall conclusion.

- ♣ 9.4.27. Consider the linear system

$$4x - y - z = 1, \quad -x + 4y - w = 2, \quad -x + 4z - w = 0, \quad -y - z + 4w = 1.$$

(a) Find the solution by using Gaussian Elimination and Back Substitution. (b) Using  $\mathbf{0}$  as your initial guess, how many iterations are required to approximate the solution to within five decimal places using (i) Jacobi iteration? (ii) Gauss-Seidel iteration? Can you estimate the spectral radii of the relevant matrices in each case? (c) Try to find the solution by using the SOR Method with the parameter  $\omega$  taking various values between .5 and 1.5. Which value of  $\omega$  gives the fastest convergence? What is the spectral radius of the SOR matrix?

- ♣ 9.4.28. (a) Find the spectral radius of the Jacobi and Gauss-Seidel iteration matrices when

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}. \quad (b) \text{ Is } A \text{ strictly diagonally dominant? (c) Use (9.76) to fix the}$$

optimal value of the SOR parameter. Verify that the spectral radius of the resulting iteration matrix agrees with the second formula in (9.76). (d) For each iterative method, predict how many iterations are needed to solve the linear system  $A\mathbf{x} = \mathbf{e}_1$  to 4 decimal places, and then verify your predictions by direct computation.

- ♣ 9.4.29. Change the matrix in Exercise 9.4.28 to  $A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 1 & 2 & -1 & 0 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$ , and answer the

same questions. Does the SOR Method with parameter given by (9.76) speed the iterations up? Why not? Can you find a value of the SOR parameter that does?

- ♣ 9.4.30. Consider the linear system  $A\mathbf{u} = \mathbf{e}_1$  in which  $A$  is the  $8 \times 8$  tridiagonal matrix with all 2's on the main diagonal and all -1's on the sub- and super-diagonals. (a) Use Exercise 8.2.47 to find the spectral radius of the Jacobi iteration method to solve  $A\mathbf{u} = \mathbf{b}$ . Does the Jacobi Method converge? (b) What is the optimal value of the SOR parameter based on (9.76)? How many Jacobi iterations are needed to match the effect of a single SOR step? (c) Test out your conclusions by using both Jacobi and SOR to approximate the solution to 3 decimal places.

- ♣ 9.4.31. How much can you speed up the convergence of the iterative solution to the pentadiagonal linear system in Exercise 9.4.22 when  $z = 4$  using SOR? Discuss.

- ◊ 9.4.32. For the matrix treated in Example 9.40, prove that (a) as  $\omega$  increases from 1 to  $8 - 4\sqrt{3}$ , the two eigenvalues move towards each other, with the larger one decreasing in magnitude; (b) if  $\omega > 8 - 4\sqrt{3}$ , the eigenvalues are complex conjugates, with larger modulus than the optimal value. (c) Can you conclude that  $\omega_* = 8 - 4\sqrt{3}$  is the optimal value for the SOR parameter?

- ♣ 9.4.33. The matrix  $A = \begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix}$  arises in the finite

difference (and finite element) discretization of the Poisson equation on a nine point square grid. Solve the linear system  $A\mathbf{u} = \mathbf{e}_5$  using (a) Gaussian Elimination; (b) Jacobi iteration; (c) Gauss-Seidel iteration; (d) SOR based on the Jacobi spectral radius.

- ♣ 9.4.34. The generalization of Exercise 9.4.33 to an  $n \times n$  grid results in an  $n^2 \times n^2$  matrix in

block tridiagonal form  $A = \begin{pmatrix} K & -I & & & \\ -I & K & -I & & \\ & -I & K & -I & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}$ , in which  $K$  is the tridiagonal

$n \times n$  matrix with 4's on the main diagonal and -1's on the sub- and super-diagonals, while  $I$  denotes the  $n \times n$  identity matrix. Use the known value of the Jacobi spectral radius  $\rho_J = \cos \frac{\pi}{n+1}$ , [86], to design an SOR Method to solve the linear system  $A\mathbf{u} = \mathbf{f}$ .

Run your method on the cases  $n = 5$  and  $\mathbf{f} = \mathbf{e}_{13}$  and  $n = 25$  and  $\mathbf{f} = \mathbf{e}_{313}$  corresponding to a unit force at the center of the grid. How much faster is the convergence rate of SOR than Jacobi and Gauss-Seidel?

- ♡ 9.4.35. If  $\mathbf{u}^{(k)}$  is an approximation to the solution to  $A\mathbf{u} = \mathbf{b}$ , then the *residual vector*  $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{u}^{(k)}$  measures how accurately the approximation solves the system.

- Show that the Jacobi iteration can be written in the form  $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + D^{-1}\mathbf{r}^{(k)}$ .
- Show that the Gauss-Seidel iteration has the form  $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + (L + D)^{-1}\mathbf{r}^{(k)}$ .
- Show that the SOR iteration has the form  $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + (\omega L + D)^{-1}\mathbf{r}^{(k)}$ .
- If  $\|\mathbf{r}^{(k)}\|$  is small, does this mean that  $\mathbf{u}^{(k)}$  is close to the solution? Explain your answer and illustrate with a couple of examples.

- 9.4.36. Let  $K$  be a positive definite  $n \times n$  matrix with eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n > 0$ . For what values of  $\varepsilon$  does the iterative system  $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \varepsilon \mathbf{r}^{(k)}$ , where  $\mathbf{r}^{(k)} = \mathbf{f} - K\mathbf{u}^{(k)}$  is the current residual vector, converge to the solution to the linear system  $K\mathbf{u} = \mathbf{f}$ ? What is the optimal value of  $\varepsilon$ , and what is the convergence rate?

## 9.5 Numerical Computation of Eigenvalues

The importance of the eigenvalues of a square matrix in a broad range of applications is amply demonstrated in this chapter and its successor. However, finding the eigenvalues and associated eigenvectors is not such an easy task. The direct method of constructing the characteristic equation of the matrix through the determinantal formula, then solving the resulting polynomial equation for the eigenvalues, and finally producing the eigenvectors by solving the associated homogeneous linear system, is hopelessly inefficient, and fraught with numerical pitfalls. We are in need of a completely new idea if we have any hopes of designing efficient numerical approximation schemes.

In this section, we develop a few of the most basic numerical algorithms for computing eigenvalues and eigenvectors. All are iterative in nature. The most direct are based on the connections between the eigenvalues and the high powers of a matrix. A more sophisticated approach, based on the  $Q R$  factorization that we learned in Section 4.3, will be presented at the end of the section. Additional computational methods for eigenvalues will appear in the following Section 9.6.

### The Power Method

We have already noted the role played by the eigenvalues and eigenvectors in the solution to linear iterative systems. Now we are going to turn the tables, and use the iterative system as a mechanism for approximating the eigenvalues, or, more correctly, selected eigenvalues of the coefficient matrix. The simplest of the resulting computational procedures is known as the *Power Method*.

We assume, for simplicity, that  $A$  is a complete<sup>†</sup>  $n \times n$  matrix. Let  $\mathbf{v}_1, \dots, \mathbf{v}_n$  denote its eigenvector basis, and  $\lambda_1, \dots, \lambda_n$  the corresponding eigenvalues. As we have learned, the solution to the linear iterative system

$$\mathbf{v}^{(k+1)} = A\mathbf{v}^{(k)}, \quad \mathbf{v}^{(0)} = \mathbf{v}, \quad (9.77)$$

is obtained by multiplying the initial vector  $\mathbf{v}$  by the successive powers of the coefficient matrix:  $\mathbf{v}^{(k)} = A^k \mathbf{v}$ . If we write the initial vector in terms of the eigenvector basis

$$\mathbf{v} = c_1 \mathbf{v}_1 + \dots + c_n \mathbf{v}_n, \quad (9.78)$$

then the solution takes the explicit form given in Theorem 9.4, namely

$$\mathbf{v}^{(k)} = A^k \mathbf{v} = c_1 \lambda_1^k \mathbf{v}_1 + \dots + c_n \lambda_n^k \mathbf{v}_n. \quad (9.79)$$

Suppose further that  $A$  has a single *dominant real* eigenvalue,  $\lambda_1$ , that is larger than all others in magnitude, so

$$|\lambda_1| > |\lambda_j| \quad \text{for all } j > 1. \quad (9.80)$$

As its name implies, this eigenvalue will eventually dominate the iteration (9.79). Indeed, since

$$|\lambda_1|^k \gg |\lambda_j|^k \quad \text{for all } j > 1 \text{ and all } k \gg 0,$$

the first term in the iterative formula (9.79) will eventually be much larger than the rest, and so, provided  $c_1 \neq 0$ ,

$$\mathbf{v}^{(k)} \simeq c_1 \lambda_1^k \mathbf{v}_1 \quad \text{for } k \gg 0.$$

Therefore, the solution to the iterative system (9.77) will, almost always, end up being a multiple of the dominant eigenvector of the coefficient matrix.

To compute the corresponding eigenvalue, we note that the  $i^{\text{th}}$  entry of the iterate  $\mathbf{v}^{(k)}$  is approximated by  $v_i^{(k)} \simeq c_1 \lambda_1^k v_{1,i}$ , where  $v_{1,i}$  is the  $i^{\text{th}}$  entry of the eigenvector  $\mathbf{v}_1$ . Thus, as long as  $v_{1,i} \neq 0$ , we can recover the dominant eigenvalue by taking a ratio between selected components of successive iterates:

$$\lambda_1 \simeq \frac{v_i^{(k)}}{v_i^{(k-1)}}, \quad \text{provided that } v_i^{(k-1)} \neq 0. \quad (9.81)$$

**Example 9.41.** Consider the matrix  $A = \begin{pmatrix} -1 & 2 & 2 \\ -1 & -4 & -2 \\ -3 & 9 & 7 \end{pmatrix}$ . As you can check, its eigenvalues and eigenvectors are

$$\lambda_1 = 3, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}, \quad \lambda_2 = -2, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \quad \lambda_3 = 1, \quad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix}.$$

Repeatedly multiplying the initial vector  $\mathbf{v} = (1, 0, 0)^T$  by  $A$  results in the iterates  $\mathbf{v}^{(k)} = A^k \mathbf{v}$  listed in the accompanying table. The last column indicates the ratio  $\lambda^{(k)} = v_1^{(k)} / v_1^{(k-1)}$  between the first components of successive iterates. (One could equally

<sup>†</sup> This is not a very severe restriction. Most matrices are complete. Moreover, perturbations caused by round-off and/or numerical inaccuracies will almost invariably make an incomplete matrix complete.

$k$	$\mathbf{v}^{(k)}$			$\lambda^{(k)}$
0	1	0	0	
1	-1	-1	-3	-1.
2	-7	11	-27	7.
3	-25	17	-69	3.5714
4	-79	95	-255	3.1600
5	-241	209	-693	3.0506
6	-727	791	-2247	3.0166
7	-2185	2057	-6429	3.0055
8	-6559	6815	-19935	3.0018
9	-19681	19169	-58533	3.0006
10	-59047	60071	-178167	3.0002
11	-177145	175097	-529389	3.0001
12	-531439	535535	-1598415	3.0000

well use the second or third components.) The ratios are converging to the dominant eigenvalue  $\lambda_1 = 3$ , while the vectors  $\mathbf{v}^{(k)}$  are converging to a very large multiple of the corresponding eigenvector  $\mathbf{v}_1 = (1, -1, 3)^T$ .

The success of the Power Method lies in the assumption that  $A$  has a unique dominant eigenvalue of maximal modulus, which, by definition, equals its spectral radius:  $|\lambda_1| = \rho(A)$ . The rate of convergence of the method is governed by the ratio  $|\lambda_2/\lambda_1|$  between the subdominant and dominant eigenvalues. Thus, the farther the dominant eigenvalue lies away from the rest, the faster the Power Method converges. We also assumed that the initial vector  $\mathbf{v}^{(0)}$  includes a nonzero multiple of the dominant eigenvector, i.e.,  $c_1 \neq 0$ . As we do not know the eigenvectors, it is not so easy to guarantee this in advance, although one must be quite unlucky to make such a poor choice of initial vector. (Of course, the stupid choice  $\mathbf{v}^{(0)} = \mathbf{0}$  is not counted.) Moreover, even if  $c_1$  happens to be 0 initially, numerical round-off error will typically come to one's rescue, since it will almost inevitably introduce a tiny component of the eigenvector  $\mathbf{v}_1$  into some iterate, and this component will eventually dominate the computation. The trick is to wait long enough for it to appear!

Since the iterates of  $A$  are, typically, getting either very large — when  $\rho(A) > 1$  — or very small — when  $\rho(A) < 1$  — the iterated vectors will be increasingly subject to numerical overflow or underflow, and the method may break down before a reasonable approximation is achieved. One way to avoid this outcome is to restrict our attention to unit vectors relative to a given norm, e.g., the Euclidean norm or the  $\infty$  norm, since their entries cannot be too large, and so are less likely to cause numerical errors in the computations. As usual, the unit vector  $\mathbf{u}^{(k)} = \|\mathbf{v}^{(k)}\|^{-1} \mathbf{v}^{(k)}$  is obtained by dividing the iterate by its norm; it can be computed directly by the modified iterative algorithm

$$\mathbf{u}^{(0)} = \frac{\mathbf{v}^{(0)}}{\|\mathbf{v}^{(0)}\|}, \quad \text{and} \quad \mathbf{u}^{(k+1)} = \frac{A\mathbf{u}^{(k)}}{\|A\mathbf{u}^{(k)}\|}. \quad (9.82)$$

If the dominant eigenvalue is positive,  $\lambda_1 > 0$ , then  $\mathbf{u}^{(k)} \rightarrow \mathbf{u}_1$  will converge to one of the

$k$	$\mathbf{u}^{(k)}$			$\lambda$
0	1	0	0	
1	-.3015	-.3015	-.9045	-1.0000
2	-.2335	.3669	-.9005	7.0000
3	-.3319	.2257	-.9159	3.5714
4	-.2788	.3353	-.8999	3.1600
5	-.3159	.2740	-.9084	3.0506
6	-.2919	.3176	-.9022	3.0166
7	-.3080	.2899	-.9061	3.0055
8	-.2973	.3089	-.9035	3.0018
9	-.3044	.2965	-.9052	3.0006
10	-.2996	.3048	-.9041	3.0002
11	-.3028	.2993	-.9048	3.0001
12	-.3007	.3030	-.9043	3.0000

two dominant unit eigenvectors (the other is  $-\mathbf{u}_1$ ). If  $\lambda_1 < 0$ , then the iterates will switch back and forth between the two eigenvectors, so  $\mathbf{u}^{(k)} \simeq \pm \mathbf{u}_1$ . In either case, the dominant eigenvalue  $\lambda_1$  is obtained as a limiting ratio between nonzero entries of  $A\mathbf{u}^{(k)}$  and  $\mathbf{u}^{(k)}$ . If some other sort of behavior is observed, it means that one of our assumptions is not valid; either  $A$  has more than one dominant eigenvalue of maximum modulus, e.g., it has a complex conjugate pair of eigenvalues of largest modulus, or it is not complete. In such cases, one can apply the more general long term behavior described in Exercise 9.2.8 to pin down the dominant eigenvalues.

**Example 9.42.** For the matrix considered in Example 9.41, starting the iterative system (9.82) with  $\mathbf{u}^{(k)} = (1, 0, 0)^T$ , the resulting unit vectors are tabulated above. The last column, being the ratio between the first components of  $A\mathbf{u}^{(k-1)}$  and  $\mathbf{u}^{(k-1)}$ , again converges to the dominant eigenvalue  $\lambda_1 = 3$ .

Variants of the Power Method for computing the other eigenvalues of the matrix are explored in the exercises.

**Remark.** See Wilkinson, [90; Chapter 2] for the perturbation theory of eigenvalues, i.e., how they can behave under small perturbations of the matrix. Wilkinson defines a *spectral condition number* to equal the product of the norms of the matrix used to place the matrix in Jordan canonical form and its inverse. The larger the spectral condition number, the more the eigenvalues deviate under perturbation. In particular symmetric matrices have spectral condition number = 1, and so their eigenvalues are well behaved under perturbations. He also gives examples of highly ill-conditioned matrices. Similarly, in [69; Section 3.3], Saad defines a condition number for an individual simple eigenvalue, and proves that it is the reciprocal of the cosine of the angle between its eigenvectors and co-eigenvectors (left eigenvectors).

---

## Exercises

♠ 9.5.1. Use the Power Method to find the dominant eigenvalue and associated eigenvector of the following matrices:

$$(a) \begin{pmatrix} -1 & -2 \\ 3 & 4 \end{pmatrix}, \quad (b) \begin{pmatrix} -5 & 2 \\ -3 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}, \quad (d) \begin{pmatrix} -2 & 0 & 1 \\ -3 & -2 & 0 \\ -2 & 5 & 4 \end{pmatrix},$$

$$(e) \begin{pmatrix} -1 & -2 & -2 \\ 1 & 2 & 5 \\ -1 & 4 & 0 \end{pmatrix}, \quad (f) \begin{pmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 2 & 2 & 2 \end{pmatrix}, \quad (g) \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad (h) \begin{pmatrix} 4 & 1 & 0 & 1 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 1 & 0 & 1 & 4 \end{pmatrix}.$$

♠ 9.5.2. Use the Power Method to find the largest singular value of the following matrices:

$$(a) \begin{pmatrix} 1 & 2 \\ -1 & 3 \end{pmatrix}, \quad (b) \begin{pmatrix} 2 & 1 & -1 \\ -2 & 3 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 2 & 2 & 1 & -1 \\ 1 & -2 & 0 & 1 \end{pmatrix}, \quad (d) \begin{pmatrix} 3 & 1 & -1 \\ 1 & -2 & 2 \\ 2 & -1 & 1 \end{pmatrix}.$$

♠ 9.5.3. Let  $T_n$  be the tridiagonal matrix whose diagonal entries are all equal to 2 and whose sub- and super-diagonal entries all equal 1. Use the Power Method to find the dominant eigenvalue of  $T_n$  for  $n = 10, 20, 50$ . Do your values agree with those in Exercise 8.2.47? How many iterations do you require to obtain 4 decimal place accuracy?

◇ 9.5.4. Prove that, for the iterative method (9.82),  $\|A\mathbf{u}^{(k)}\| \rightarrow |\lambda_1|$ . Assuming  $\lambda_1$  is real, explain how to deduce its sign.

◇ 9.5.5. *The Inverse Power Method.* Let  $A$  be a nonsingular matrix. (a) Show that the eigenvalues of  $A^{-1}$  are the reciprocals  $1/\lambda$  of the eigenvalues of  $A$ . How are the eigenvectors related? (b) Show how to use the Power Method on  $A^{-1}$  to produce the smallest (in modulus) eigenvalue of  $A$ . (c) What is the rate of convergence of the algorithm? (d) Design a practical iterative algorithm based on the (permuted)  $LU$  decomposition of  $A$ .

♠ 9.5.6. Apply the Inverse Power Method of Exercise 9.5.7 to the find the smallest eigenvalue of the matrices in Exercise 9.5.1.

◇ 9.5.7. *The Shifted Inverse Power Method.* Suppose that  $\mu$  is *not* an eigenvalue of  $A$ .

(a) Show that the iterative system  $\mathbf{u}^{(k+1)} = (A - \mu I)^{-1}\mathbf{u}^{(k)}$  converges to the eigenvector of  $A$  corresponding to the eigenvalue  $\lambda^*$  that is *closest* to  $\mu$ . Explain how to find the eigenvalue  $\lambda^*$ . (b) What is the rate of convergence of the algorithm? (c) What happens if  $\mu$  is an eigenvalue?

♠ 9.5.8. Apply the Shifted Inverse Power Method of Exercise 9.5.7 to the find the eigenvalue closest to  $\mu = .5$  of the matrices in Exercise 9.5.1.

9.5.9. Suppose that  $A\mathbf{u}^{(k)} = \mathbf{0}$  in the iterative procedure (9.82). What does this indicate?

♠ 9.5.10. (i) Explain how to use the Deflation Method of Exercise 8.2.51 to find the subdominant eigenvalue of a nonsingular matrix  $A$ . (ii) Apply your method to the matrices listed in Exercise 9.5.1.

---

## The QR Algorithm

As stated, the Power Method produces only the dominant (largest in magnitude) eigenvalue of a matrix  $A$ . The Inverse Power Method of Exercise 9.5.5 can be used to find the smallest eigenvalue. Additional eigenvalues can be found by using the Shifted Inverse Power Method of Exercise 9.5.7, or the Deflation Method of Exercise 9.5.10. However, if we need to know

all the eigenvalues, such piecemeal methods are too time-consuming to be of much practical value.

The most popular scheme for simultaneously approximating all the eigenvalues of a matrix  $A$  is the remarkable  $QR$  algorithm, first proposed in 1961 by John Francis, [29], and Vera Kublanovskaya, [51]. The underlying idea is simple, but surprising. The first step is to factor the matrix

$$A = A_0 = Q_0 R_0$$

into a product of an orthogonal matrix  $Q_0$  and a positive (i.e., with all positive entries along the diagonal) upper triangular matrix  $R_0$  by using the Gram–Schmidt orthogonalization procedure of Theorem 4.24, or, even better, the numerically stable version described in (4.28). Next, multiply the two factors together *in the wrong order!* The result is the new matrix

$$A_1 = R_0 Q_0.$$

We then repeat these two steps. Thus, we next factor

$$A_1 = Q_1 R_1$$

using the Gram–Schmidt process, and then multiply the factors in the reverse order to produce

$$A_2 = R_1 Q_1.$$

The complete algorithm can be written as

$$A = A_0 = Q_0 R_0, \quad A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1}, \quad k = 0, 1, 2, \dots, \quad (9.83)$$

where  $Q_k, R_k$  come from the previous step, and the subsequent orthogonal matrix  $Q_{k+1}$  and positive upper triangular matrix  $R_{k+1}$  are computed directly from  $A_{k+1} = R_k Q_k$  by applying the numerically stable form of the Gram–Schmidt algorithm.

The astonishing fact is that, for many matrices  $A$  with all real eigenvalues, the iterates  $A_k \rightarrow V$  converge to an upper triangular matrix  $V$  whose diagonal entries are the eigenvalues of  $A$ . Thus, after a sufficient number of iterations, say  $m$ , the matrix  $A_m$  will have very small entries below the diagonal, and one can read off a complete system of (approximate) eigenvalues along its diagonal. For each eigenvalue, the computation of the corresponding eigenvector can be most efficiently accomplished by applying the Shifted Inverse Power Method of Exercise 9.5.7 with parameter  $\mu$  chosen near the computed eigenvalue.

**Example 9.43.** Consider the matrix  $A = \begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix}$ . The initial Gram–Schmidt factorization  $A = Q_0 R_0$  yields

$$Q_0 \simeq \begin{pmatrix} .7071 & -.7071 \\ .7071 & .7071 \end{pmatrix}, \quad R_0 \simeq \begin{pmatrix} 2.8284 & 2.8284 \\ 0 & 1.4142 \end{pmatrix}.$$

These are multiplied in the reverse order to give

$$A_1 = R_0 Q_0 = \begin{pmatrix} 4 & 0 \\ 1 & 1 \end{pmatrix}.$$

We refactor  $A_1 = Q_1 R_1$  via Gram–Schmidt, and then reverse multiply to produce

$$Q_1 \simeq \begin{pmatrix} .9701 & -.2425 \\ .2425 & .9701 \end{pmatrix}, \quad R_1 \simeq \begin{pmatrix} 4.1231 & .2425 \\ 0 & .9701 \end{pmatrix},$$

$$A_2 = R_1 Q_1 \simeq \begin{pmatrix} 4.0588 & -.7647 \\ .2353 & .9412 \end{pmatrix}.$$

The next iteration yields

$$Q_2 \simeq \begin{pmatrix} .9983 & -.0579 \\ .0579 & .9983 \end{pmatrix}, \quad R_2 \simeq \begin{pmatrix} 4.0656 & -.7090 \\ 0 & .9839 \end{pmatrix},$$

$$A_3 = R_2 Q_2 \simeq \begin{pmatrix} 4.0178 & -.9431 \\ .0569 & .9822 \end{pmatrix}.$$

Continuing in this manner, after 9 iterations we obtain, to four decimal places,

$$Q_9 \simeq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R_9 \simeq \begin{pmatrix} 4 & -1 \\ 0 & 1 \end{pmatrix}, \quad A_{10} = R_9 Q_9 \simeq \begin{pmatrix} 4 & -1 \\ 0 & 1 \end{pmatrix}.$$

The eigenvalues of  $A$ , namely 4 and 1, appear along the diagonal of  $A_{10}$ . Additional iterations produce very little further change, although they can be used for increasing the numerical accuracy of the computed eigenvalues.

If the original matrix  $A$  happens to be symmetric and positive definite, then the limiting matrix  $A_k \rightarrow V = \Lambda$  is, in fact, the diagonal matrix containing the eigenvalues of  $A$ . Moreover, if, in this case, we recursively define

$$S_k = S_{k-1} Q_k = Q_0 Q_1 \cdots Q_{k-1} Q_k, \quad (9.84)$$

which then have, as their limit,  $S_k \rightarrow S$ , an orthogonal matrix, whose columns are the orthonormal eigenvector basis of  $A$ .

**Example 9.44.** Consider the symmetric matrix  $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & -1 \\ 0 & -1 & 6 \end{pmatrix}$ . The initial  $A = Q_0 R_0$  factorization produces

$$S_0 = Q_0 \simeq \begin{pmatrix} .8944 & -.4082 & -.1826 \\ .4472 & .8165 & .3651 \\ 0 & -.4082 & .9129 \end{pmatrix}, \quad R_0 \simeq \begin{pmatrix} 2.2361 & 2.2361 & -.4472 \\ 0 & 2.4495 & -3.2660 \\ 0 & 0 & 5.1121 \end{pmatrix},$$

and so

$$A_1 = R_0 Q_0 \simeq \begin{pmatrix} 3.0000 & 1.0954 & 0 \\ 1.0954 & 3.3333 & -2.0870 \\ 0 & -2.0870 & 4.6667 \end{pmatrix}.$$

We refactor  $A_1 = Q_1 R_1$  and reverse multiply to produce

$$Q_1 \simeq \begin{pmatrix} .9393 & -.2734 & -.2071 \\ .3430 & .7488 & .5672 \\ 0 & -.6038 & .7972 \end{pmatrix}, \quad S_1 = S_0 Q_1 \simeq \begin{pmatrix} .7001 & -.4400 & -.5623 \\ .7001 & .2686 & .6615 \\ -.1400 & -.8569 & .4962 \end{pmatrix},$$

$$R_1 \simeq \begin{pmatrix} 3.1937 & 2.1723 & -.7158 \\ 0 & 3.4565 & -4.3804 \\ 0 & 0 & 2.5364 \end{pmatrix}, \quad A_2 = R_1 Q_1 \simeq \begin{pmatrix} 3.7451 & 1.1856 & 0 \\ 1.1856 & 5.2330 & -1.5314 \\ 0 & -1.5314 & 2.0219 \end{pmatrix}.$$

Continuing in this manner, after 10 iterations we have

$$Q_{10} \simeq \begin{pmatrix} 1.0000 & -.0067 & 0 \\ .0067 & 1.0000 & .0001 \\ 0 & -.0001 & 1.0000 \end{pmatrix}, \quad S_{10} \simeq \begin{pmatrix} .0753 & -.5667 & -.8205 \\ .3128 & -.7679 & .5591 \\ -.9468 & -.2987 & .1194 \end{pmatrix},$$

$$R_{10} \simeq \begin{pmatrix} 6.3229 & .0647 & 0 \\ 0 & 3.3582 & -.0006 \\ 0 & 0 & 1.3187 \end{pmatrix}, \quad A_{11} \simeq \begin{pmatrix} 6.3232 & .0224 & 0 \\ .0224 & 3.3581 & -.0002 \\ 0 & -.0002 & 1.3187 \end{pmatrix}.$$

After 20 iterations, the process has completely settled down, and

$$Q_{20} \simeq \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad S_{20} \simeq \begin{pmatrix} .0710 & -.5672 & -.8205 \\ .3069 & -.7702 & .5590 \\ -.9491 & -.2915 & .1194 \end{pmatrix},$$

$$R_{20} \simeq \begin{pmatrix} 6.3234 & .0001 & 0 \\ 0 & 3.3579 & 0 \\ 0 & 0 & 1.3187 \end{pmatrix}, \quad A_{21} \simeq \begin{pmatrix} 6.3234 & 0 & 0 \\ 0 & 3.3579 & 0 \\ 0 & 0 & 1.3187 \end{pmatrix}.$$

The eigenvalues of  $A$  appear along the diagonal of  $A_{21}$ , while the columns of  $S_{20}$  are the corresponding orthonormal eigenvector basis, listed in the same order as the eigenvalues, both correct to 4 decimal places.

We will devote the remainder of this section to a justification of the  $QR$  algorithm for a class of matrices. We will assume that  $A$  is symmetric, and that its (necessarily real) eigenvalues satisfy

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0. \quad (9.85)$$

According to the Spectral Theorem 8.38, the corresponding unit eigenvectors  $\mathbf{u}_1, \dots, \mathbf{u}_n$  (in the Euclidean norm) form an orthonormal basis of  $\mathbb{R}^n$ . Our analysis can be adapted to a broader class of matrices, but this will suffice to expose the main ideas without unduly complicating the exposition.

The secret is that the  $QR$  algorithm is, in fact, a well-disguised adaptation of the more primitive Power Method. If we were to use the Power Method to capture all the eigenvectors and eigenvalues of  $A$ , the first thought might be to try to perform it simultaneously on a complete basis  $\mathbf{v}_1^{(0)}, \dots, \mathbf{v}_n^{(0)}$  of  $\mathbb{R}^n$  instead of just one individual vector. The problem is that, for almost all vectors, the power iterates  $\mathbf{v}_j^{(k)} = A^k \mathbf{v}_j^{(0)}$  all tend to a multiple of the dominant eigenvector  $\mathbf{u}_1$ . Normalizing the vectors at each step, as in (9.82), is not any better, since then they merely converge to one of the two dominant unit eigenvectors  $\pm \mathbf{u}_1$ . However, if, inspired by the form of the eigenvector basis, we *orthonormalize* the vectors at each step, then we effectively prevent them from all accumulating at the same dominant unit eigenvector, and so, with some luck, the resulting vectors will converge to the full system of eigenvectors. Since orthonormalizing a basis via the Gram–Schmidt process is equivalent to a  $QR$  matrix factorization, the mechanics of the algorithm becomes less surprising.

In detail, we start with any orthonormal basis, which, for simplicity, we take to be the standard basis vectors of  $\mathbb{R}^n$ , and so  $\mathbf{u}_1^{(0)} = \mathbf{e}_1, \dots, \mathbf{u}_n^{(0)} = \mathbf{e}_n$ . At the  $k^{\text{th}}$  stage of the algorithm, we set  $\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_n^{(k)}$  to be the orthonormal vectors that result from applying the Gram–Schmidt algorithm to the power vectors  $\mathbf{v}_j^{(k)} = A^k \mathbf{e}_j$ . In matrix language, the vectors  $\mathbf{v}_1^{(k)}, \dots, \mathbf{v}_n^{(k)}$  are merely the columns of  $A^k$ , and the orthonormal basis  $\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_n^{(k)}$  are the columns of the orthogonal matrix  $S_k$  in the  $QR$  decomposition of the  $k^{\text{th}}$  power of  $A$ , which we denote by

$$A^k = S_k P_k, \quad (9.86)$$

where  $P_k$  is positive upper triangular, meaning all its diagonal entries are positive. Note that, in view of (9.83)

$$A = Q_0 R_0, \quad A^2 = Q_0 R_0 Q_0 R_0 = Q_0 Q_1 R_1 R_0,$$

$$A^3 = Q_0 R_0 Q_0 R_0 Q_0 R_0 = Q_0 Q_1 R_1 Q_1 R_1 R_0 = Q_0 Q_1 Q_2 R_2 R_1 R_0,$$

and, in general,

$$A^k = (Q_0 Q_1 \cdots Q_{k-1}) (R_{k-1} \cdots R_1 R_0). \quad (9.87)$$

Proposition 4.23 tells us that the product of orthogonal matrices is also orthogonal. The product of positive upper triangular matrices is also positive upper triangular. Therefore, comparing (9.86, 87) and invoking the uniqueness of the  $QR$  factorization, we conclude that

$$S_k = Q_0 Q_1 \cdots Q_{k-1} = S_{k-1} Q_{k-1}, \quad P_k = R_{k-1} \cdots R_1 R_0 = R_{k-1} P_{k-1}. \quad (9.88)$$

Let  $S = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n)$  denote an orthogonal matrix whose columns are unit eigenvectors of  $A$ . The Spectral Theorem 8.38 tells us that

$$A = S \Lambda S^T, \quad \text{where} \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$$

is the diagonal eigenvalue matrix. Substituting the spectral factorization into (9.86) yields

$$A^k = S \Lambda^k S^T = S_k P_k.$$

We now make one additional assumption on the matrix  $A$  by requiring that  $S^T$  be a regular matrix, meaning that it can be factored,  $S^T = LU$ , as the product of a lower unitriangular matrix and an upper triangular matrix. We can further assume, without loss of generality, that the diagonal entries of  $U$  — that is, the pivots of  $S^T$  — are all positive. Indeed, by Exercise 1.3.31, this can be arranged by multiplying each row of  $S^T$  by the sign of its pivot, which amounts to possibly replacing some of the unit eigenvectors  $\mathbf{u}_j$  by their negatives  $-\mathbf{u}_j$ , which is allowed, since it does not affect their status as an orthonormal eigenvector basis. Regularity of  $S^T$  holds generically, and is the analogue of the condition that our initial vector in the Power Method includes a nonzero component of the dominant eigenvector.

Under these two assumptions,

$$A^k = S \Lambda^k LU = S_k P_k, \quad \text{and hence} \quad S \Lambda^k L = S_k P_k U^{-1}.$$

Multiplying on the right by  $\Lambda^{-k}$ , we obtain

$$S \Lambda^k L \Lambda^{-k} = S_k T_k, \quad \text{where} \quad T_k = P_k U^{-1} \Lambda^{-k} \quad (9.89)$$

is also a positive upper triangular matrix, since  $P_k, U, \Lambda$  are all of that form.

Now consider what happens as  $k \rightarrow \infty$ . The entries of the lower triangular matrix  $N = \Lambda^k L \Lambda^{-k}$  are

$$n_{ij} = \begin{cases} l_{ij}(\lambda_i/\lambda_j)^k, & i > j, \\ l_{ii} = 1, & i = j, \\ 0, & i < j. \end{cases}$$

Since we are assuming  $|\lambda_i| < |\lambda_j|$  when  $i > j$ , we immediately deduce that

$$\Lambda^k L \Lambda^{-k} \rightarrow \mathbf{I}, \quad \text{and hence} \quad S_k T_k = S \Lambda^k L \Lambda^{-k} \rightarrow S \quad \text{as} \quad k \rightarrow \infty.$$

We now appeal to the following lemma, whose proof will be given after we finish the justification of the  $QR$  algorithm.

**Lemma 9.45.** Let  $S_1, S_2, \dots$  and  $S$  be orthogonal matrices and  $T_1, T_2, \dots$  positive upper triangular matrices. Then  $S_k T_k \rightarrow S$  as  $k \rightarrow \infty$  if and only if  $S_k \rightarrow S$  and  $T_k \rightarrow \mathbf{I}$ .

Lemma 9.45 implies that, as claimed, the orthogonal matrices  $S_k$  do converge to the orthogonal eigenvector matrix  $S$ . Moreover, by (9.88–89),

$$R_k = P_k P_{k-1}^{-1} = (T_k \Lambda^k U^{-1}) (T_{k-1} \Lambda^{k-1} U^{-1})^{-1} = T_k \Lambda T_{k-1}^{-1}.$$

Since both  $T_k$  and  $T_{k-1}$  converge to the identity matrix,  $R_k$  converges to the diagonal eigenvalue matrix  $\Lambda$ , as claimed. The eigenvalues appear in decreasing order along the diagonal — this is a consequence of our regularity assumption on the transposed eigenvector matrix  $S^T$ .

**Theorem 9.46.** If  $A$  is positive definite with all simple eigenvalues, and its transposed eigenvector matrix  $S^T$  is regular, then the matrices  $S_k \rightarrow S$  and  $R_k \rightarrow \Lambda$  appearing in the  $QR$  algorithm applied to  $A$  converge to, respectively, the eigenvector matrix  $S$  and the diagonal eigenvalue matrix  $\Lambda$ .

**Remark.** If  $A$  is symmetric and has all simple eigenvalues, then, for suitably large  $\alpha \gg 0$ , the *shifted matrix*  $\tilde{A} = A + \alpha I$  is positive definite, has the same eigenvectors as  $A$ , and has simple shifted eigenvalues  $\tilde{\lambda}_k = \lambda_k + \alpha$ . Thus, one can run the  $QR$  algorithm to determine the eigenvalues and eigenvectors of  $\tilde{A}$ , and hence those of  $A$  by undoing the shift.

The last remaining item is a proof of Lemma 9.45. We write

$$S = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n), \quad S_k = \left( \mathbf{u}_1^{(k)} \ \mathbf{u}_2^{(k)} \ \dots \ \mathbf{u}_n^{(k)} \right),$$

in columnar form. Let  $t_{ij}^{(k)}$  denote the entries of the positive upper triangular matrix  $T_k$ . The last column of the limiting equation  $S_k T_k \rightarrow S$  reads  $t_{nn}^{(k)} \mathbf{u}_n^{(k)} \rightarrow \mathbf{u}_n$ . Since both  $\mathbf{u}_n^{(k)}$  and  $\mathbf{u}_n$  are unit vectors, and  $t_{nn}^{(k)} > 0$ , it follows that

$$\|t_{nn}^{(k)} \mathbf{u}_n^{(k)}\| = t_{nn}^{(k)} \rightarrow \| \mathbf{u}_n \| = 1, \quad \text{and hence the last column } \mathbf{u}_n^{(k)} \rightarrow \mathbf{u}_n.$$

The next to last column reads

$$t_{n-1,n-1}^{(k)} \mathbf{u}_{n-1}^{(k)} + t_{n-1,n}^{(k)} \mathbf{u}_n^{(k)} \rightarrow \mathbf{u}_{n-1}.$$

Taking the inner product with  $\mathbf{u}_n^{(k)} \rightarrow \mathbf{u}_n$  and using orthonormality, we deduce  $t_{n-1,n}^{(k)} \rightarrow 0$ , and so  $t_{n-1,n-1}^{(k)} \mathbf{u}_{n-1}^{(k)} \rightarrow \mathbf{u}_{n-1}$ , which, by the previous reasoning, implies  $t_{n-1,n-1}^{(k)} \rightarrow 1$  and  $\mathbf{u}_{n-1}^{(k)} \rightarrow \mathbf{u}_{n-1}$ . The proof is completed by working backwards through the remaining columns, using a similar argument at each step. The remaining details are left to the interested reader.

## Exercises

9.5.11. Apply the  $QR$  algorithm to the following symmetric matrices to find their eigenvalues

and eigenvectors to 2 decimal places: (a)  $\begin{pmatrix} 1 & 2 \\ 2 & 6 \end{pmatrix}$ , (b)  $\begin{pmatrix} 3 & -1 \\ -1 & 5 \end{pmatrix}$ , (c)  $\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 3 \\ 0 & 3 & 1 \end{pmatrix}$ ,

(d)  $\begin{pmatrix} 2 & 5 & 0 \\ 5 & 0 & -3 \\ 0 & -3 & 3 \end{pmatrix}$ , (e)  $\begin{pmatrix} 3 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 3 \end{pmatrix}$ , (f)  $\begin{pmatrix} 6 & 1 & -1 & 0 \\ 1 & 8 & 1 & -1 \\ -1 & 1 & 4 & 1 \\ 0 & -1 & 1 & 3 \end{pmatrix}$ .

9.5.12. Show that applying the  $QR$  algorithm to the matrix  $A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 7 & 2 \\ 1 & 2 & 7 \end{pmatrix}$  results in a

diagonal matrix with the eigenvalues on the diagonal, but not in decreasing order. Explain.

9.5.13. Apply the  $QR$  algorithm to the following non-symmetric matrices to find their eigenvalues to 3 decimal places:

$$(a) \begin{pmatrix} -1 & -2 \\ 3 & 4 \end{pmatrix}, (b) \begin{pmatrix} 2 & 3 \\ 1 & 5 \end{pmatrix}, (c) \begin{pmatrix} 2 & 1 & 0 \\ 2 & 0 & -3 \\ 0 & -2 & 1 \end{pmatrix}, (d) \begin{pmatrix} 2 & 5 & 1 \\ 2 & -1 & 3 \\ 4 & 5 & 3 \end{pmatrix}, (e) \begin{pmatrix} 6 & 1 & 7 & 9 \\ 6 & 8 & 14 & 9 \\ 3 & 1 & 4 & 6 \\ 3 & 2 & 5 & 3 \end{pmatrix}.$$

9.5.14. The matrix  $A = \begin{pmatrix} -1 & 2 & 1 \\ -2 & 3 & 1 \\ -2 & 2 & 2 \end{pmatrix}$  has a double eigenvalue of 1, and so our proof of

convergence of the  $QR$  algorithm doesn't apply. Does the  $QR$  algorithm find its eigenvalues?

9.5.15. Explain why the  $QR$  algorithm fails to find the eigenvalues of the matrices

$$(a) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, (b) \begin{pmatrix} -2 & 1 & 0 \\ 0 & -2 & 1 \\ 1 & 0 & -2 \end{pmatrix}, (c) \begin{pmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & -1 \end{pmatrix}.$$

◇ 9.5.16. Prove that all of the matrices  $A_k$  defined in (9.83) have the same eigenvalues.

◇ 9.5.17. (a) Prove that if  $A$  is symmetric and tridiagonal, then all matrices  $A_k$  appearing in the  $QR$  algorithm are also symmetric and tridiagonal. *Hint:* First prove symmetry.

(b) Is the result true if  $A$  is not symmetric — only tridiagonal?

## Tridiagonalization

In practical implementations, the direct  $QR$  algorithm often takes overly long before providing reasonable approximations to the eigenvalues of large matrices. Fortunately, the algorithm can be made much more efficient by a simple preprocessing step. The key observation is that the  $QR$  algorithm preserves the class of symmetric tridiagonal matrices, and, like Gaussian Elimination, is much faster when applied to this class. Moreover, by applying a sequence of Householder reflection matrices (4.35), we can convert any symmetric matrix into tridiagonal form while preserving all the eigenvalues. Thus, by first using the Householder tridiagonalization process, and then applying the  $QR$  Method to the resulting tridiagonal matrix, we obtain an efficient and practical algorithm for computing eigenvalues of large symmetric matrices. Generalizations to non-symmetric matrices will be briefly considered at the end of the section.

In Householder's approach to the  $QR$  factorization, we were able to convert the matrix  $A$  to upper triangular form  $R$  by a sequence of elementary reflection matrices. Unfortunately, this procedure does not preserve the eigenvalues of the matrix — the diagonal entries of  $R$  are *not* the eigenvalues — and so we need to be a bit more clever here. We begin by recalling, from Exercise 8.2.32, that similar matrices have the same eigenvalues (but not the same eigenvectors).

**Lemma 9.47.** If  $H = I - 2\mathbf{u}\mathbf{u}^T$  is an elementary reflection matrix, with  $\mathbf{u} \in \mathbb{R}^n$  a unit vector (under the Euclidean norm), then  $A$  and  $B = HAH$  are similar matrices and hence have the same eigenvalues.

*Proof:* It suffices to note that, according to (4.37),  $H^{-1} = H$ , and hence  $B = H^{-1}AH$  is similar to  $A$ . *Q.E.D.*

Now, starting with a symmetric  $n \times n$  matrix  $A$ , our goal is to devise a similar tridiagonal matrix by applying a sequence of Householder reflections. Using the Euclidean norm, we

begin by setting

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix}, \quad \mathbf{y}_1 = \begin{pmatrix} 0 \\ \pm r_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{where} \quad r_1 = \|\mathbf{x}_1\| = \|\mathbf{y}_1\|,$$

so that  $\mathbf{x}_1$  contains all the off-diagonal entries of the first column of  $A$ . Let

$$H_1 = \mathbf{I} - 2 \mathbf{u}_1 \mathbf{u}_1^T, \quad \text{where} \quad \mathbf{u}_1 = \frac{\mathbf{x}_1 - \mathbf{y}_1}{\|\mathbf{x}_1 - \mathbf{y}_1\|},$$

be the corresponding elementary reflection matrix that maps  $\mathbf{x}_1$  to  $\mathbf{y}_1$ . Either the plus or the minus sign in the formula for  $\mathbf{y}_1$  works in the algorithm; a good choice is to set it to be the opposite of the sign of the entry  $a_{21}$ , which helps minimize the possible effects of round-off error in computing the unit vector  $\mathbf{u}_1$ . By direct computation, based on Lemma 4.28 and the fact that the first entry of  $\mathbf{u}_1$  is zero, we obtain

$$A_2 = H_1 A H_1 = \begin{pmatrix} a_{11} & r_1 & 0 & \dots & 0 \\ r_1 & \tilde{a}_{22} & \tilde{a}_{23} & \dots & \tilde{a}_{2n} \\ 0 & \tilde{a}_{32} & \tilde{a}_{33} & \dots & \tilde{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{n2} & \tilde{a}_{n3} & \dots & \tilde{a}_{nn} \end{pmatrix} \quad (9.90)$$

for certain  $\tilde{a}_{ij}$ , whose explicit formulae are not needed. Thus, by a single Householder transformation, we convert  $A$  into a similar matrix  $A_2$  whose first row and column are in tridiagonal form. We repeat the process on the lower right  $(n-1) \times (n-1)$  submatrix of  $A_2$ . We set

$$\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ \tilde{a}_{32} \\ \tilde{a}_{42} \\ \vdots \\ \tilde{a}_{n2} \end{pmatrix}, \quad \mathbf{y}_2 = \begin{pmatrix} 0 \\ 0 \\ \pm r_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{where} \quad r_2 = \|\mathbf{x}_2\| = \|\mathbf{y}_2\|,$$

and the  $\pm$  sign is chosen to be the opposite of that of  $\tilde{a}_{32}$ . Setting

$$H_2 = \mathbf{I} - 2 \mathbf{u}_2 \mathbf{u}_2^T, \quad \text{where} \quad \mathbf{u}_2 = \frac{\mathbf{x}_2 - \mathbf{y}_2}{\|\mathbf{x}_2 - \mathbf{y}_2\|},$$

we construct the similar matrix

$$A_3 = H_2 A_2 H_2 = \begin{pmatrix} a_{11} & r_1 & 0 & 0 & \dots & 0 \\ r_1 & \tilde{a}_{22} & r_2 & 0 & \dots & 0 \\ 0 & r_2 & \hat{a}_{33} & \hat{a}_{34} & \dots & \hat{a}_{3n} \\ 0 & 0 & \hat{a}_{43} & \hat{a}_{44} & \dots & \hat{a}_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \hat{a}_{n3} & \hat{a}_{n4} & \dots & \hat{a}_{nn} \end{pmatrix},$$

whose first two rows and columns are now in tridiagonal form. The remaining steps in the algorithm should now be clear. Thus, the final result is a tridiagonal matrix  $T = A_n$  that has the *same eigenvalues* (but not the same eigenvectors) as the original symmetric matrix  $A$ . Let us illustrate the method by an example.

**Example 9.48.** To tridiagonalize  $A = \begin{pmatrix} 4 & 1 & -1 & 2 \\ 1 & 4 & 1 & -1 \\ -1 & 1 & 4 & 1 \\ 2 & -1 & 1 & 4 \end{pmatrix}$ , we begin with its first column. We set  $\mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 2 \end{pmatrix}$ , so that  $\mathbf{y}_1 = \begin{pmatrix} 0 \\ \sqrt{6} \\ 0 \\ 0 \end{pmatrix} \approx \begin{pmatrix} 0 \\ 2.4495 \\ 0 \\ 0 \end{pmatrix}$ . Therefore, the unit vector and corresponding Householder matrix are

$$\mathbf{u}_1 = \frac{\mathbf{x}_1 - \mathbf{y}_1}{\|\mathbf{x}_1 - \mathbf{y}_1\|} = \begin{pmatrix} 0 \\ .8391 \\ -.2433 \\ .4865 \end{pmatrix}, \quad H_1 = \mathbf{I} - 2\mathbf{u}_1\mathbf{u}_1^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -.4082 & .4082 & -.8165 \\ 0 & .4082 & .8816 & .2367 \\ 0 & -.8165 & .2367 & .5266 \end{pmatrix}.$$

We compute

$$A_2 = H_1 A H_1 = \begin{pmatrix} 4.0000 & -2.4495 & 0 & 0 \\ -2.4495 & 2.3333 & -.3865 & -.8599 \\ 0 & -.3865 & 4.9440 & -.1246 \\ 0 & -.8599 & -.1246 & 4.7227 \end{pmatrix}.$$

In the next phase,  $\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ -.3865 \\ -.8599 \end{pmatrix}$ ,  $\mathbf{y}_2 = \begin{pmatrix} 0 \\ 0 \\ -.9428 \\ 0 \end{pmatrix}$ , so  $\mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -.8396 \\ -.5431 \end{pmatrix}$ , and

$$H_2 = \mathbf{I} - 2\mathbf{u}_2\mathbf{u}_2^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -.4100 & -.9121 \\ 0 & 0 & -.9121 & .4100 \end{pmatrix}.$$

The resulting matrix

$$T = A_3 = H_2 A_2 H_2 = \begin{pmatrix} 4.0000 & -2.4495 & 0 & 0 \\ -2.4495 & 2.3333 & .9428 & 0 \\ 0 & .9428 & 4.6667 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}$$

is now in tridiagonal form.

Since the final tridiagonal matrix  $T$  has the same eigenvalues as  $A$ , we can apply the  $QR$  algorithm to  $T$  to approximate the common eigenvalues. According to Exercise 9.5.17, if  $A = A_1$  is tridiagonal, so are all its  $QR$  iterates  $A_2, A_3, \dots$ . Moreover, far fewer arithmetic operations are required; in Exercise 9.5.25, you are asked to quantify this. For instance, in the preceding example, after we apply 20 iterations of the  $QR$  algorithm directly to  $T$ , the upper triangular factor has become

$$R_{20} = \begin{pmatrix} 6.0000 & -.0065 & 0 & 0 \\ 0 & 4.5616 & 0 & 0 \\ 0 & 0 & 5.0000 & 0 \\ 0 & 0 & 0 & .4384 \end{pmatrix}.$$

The eigenvalues of  $T$ , and hence also of  $A$ , appear along the diagonal, and are correct to 4 decimal places. As noted earlier, with the eigenvalues in hand the corresponding eigenvectors can then be found via the Shifted Inverse Power Method of Exercise 9.5.7.

Finally, even if  $A$  is not symmetric, one can still apply the same sequence of Householder reflections to simplify it. The final result is no longer tridiagonal, but rather a similar *upper Hessenberg matrix*, which means that all entries below its subdiagonal are zero, but those above its superdiagonal are not necessarily zero. For instance, a  $5 \times 5$  upper Hessenberg matrix looks like

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix},$$

where the starred entries can be anything. It can be proved that the  $QR$  algorithm maintains the upper Hessenberg form, and, while not as efficient as in the tridiagonal case, still yields a significant savings in computational effort required to find the common eigenvalues.

If  $A$  has no eigenvalues of the same magnitude, which, in particular, requires all its eigenvalues to be simple, then application of the tridiagonal  $QR$  algorithm to its tridiagonalization will, usually, produce its eigenvalues. More generally, if  $A$  has  $k$  eigenvalues of the same magnitude, then the  $QR$  algorithm, applied either directly to  $A$ , or to its tridiagonalization, will, again generically, converge to a block upper triangular matrix, with an  $k \times k$  matrix in the block diagonal slot that has these same eigenvalues. Thus, for example, if  $A$  is a real matrix with simple real and complex eigenvalues, then each complex conjugate pair will be the eigenvalues of one of the  $2 \times 2$  matrices appearing on the diagonal of the eventual  $QR$  iterates, while the real eigenvalues will appear directly (in a  $1 \times 1$  “block”) on the diagonal.

Further details and results can be found in [21, 66, 69, 89, 90].

## Exercises

9.5.18. Use Householder matrices to convert the following matrices into tridiagonal form:

$$(a) \begin{pmatrix} 8 & -7 & 2 \\ -7 & 17 & -7 \\ 2 & -7 & 8 \end{pmatrix}, \quad (b) \begin{pmatrix} 5 & 1 & -2 & 1 \\ 1 & 5 & 1 & -2 \\ -2 & 1 & 5 & 1 \\ 1 & -2 & 1 & 5 \end{pmatrix}, \quad (c) \begin{pmatrix} 4 & 0 & -1 & 1 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 2 & 0 \\ 1 & -1 & 0 & 3 \end{pmatrix}.$$

♠ 9.5.19. Find the eigenvalues, to 2 decimal places, of the matrices in Exercise 9.5.18 by applying the  $QR$  algorithm to the tridiagonal form.

♠ 9.5.20. Use the tridiagonal  $QR$  Method to find the singular values of  $A = \begin{pmatrix} 2 & 2 & 1 & -1 \\ 1 & -2 & 0 & 1 \\ 0 & -1 & 2 & 2 \end{pmatrix}$ .

9.5.21. Use Householder matrices to convert the following matrices into upper Hessenberg form:

$$(a) \begin{pmatrix} 3 & -1 & 2 \\ 1 & 3 & -4 \\ 2 & -1 & -1 \end{pmatrix}, \quad (b) \begin{pmatrix} 3 & 2 & -1 & 1 \\ 2 & 4 & 0 & 1 \\ 0 & 1 & 2 & -6 \\ 1 & 0 & -5 & 1 \end{pmatrix}, \quad (c) \begin{pmatrix} 1 & 0 & -1 & 1 \\ 2 & 1 & 1 & -1 \\ -1 & 0 & 1 & 3 \\ 3 & -1 & 1 & 4 \end{pmatrix}.$$

♠ 9.5.22. Find the eigenvalues, to 2 decimal places, of the matrices in Exercise 9.5.21 by applying the  $QR$  algorithm to the upper Hessenberg form.

9.5.23. Prove that the effect of the first Householder reflection is as given in (9.90).

9.5.24. What is the effect of tridiagonalization on the eigenvectors of the matrix?

- ◇ 9.5.25. (a) How many arithmetic operations — multiplications/divisions and additions/subtractions — are required to place a generic  $n \times n$  symmetric matrix into tridiagonal form? (b) How many operations are needed to perform one iteration of the  $QR$  algorithm on an  $n \times n$  tridiagonal matrix? (c) How much faster, on average, is the tridiagonal algorithm than the direct  $QR$  algorithm for finding the eigenvalues of a symmetric matrix?
- 9.5.26. Write out a pseudocode program to tridiagonalize a matrix. The input should be an  $n \times n$  matrix  $A$ , and the output should be the Householder unit vectors  $\mathbf{u}_1, \dots, \mathbf{u}_{n-1}$  and the tridiagonal matrix  $R$ . Does your program produce the upper Hessenberg form when the input matrix is not symmetric?
- ◇ 9.5.27. Prove that in the  $H = LU$  factorization of a regular upper Hessenberg matrix, the lower triangular factor  $L$  is bidiagonal, as in (1.67).

## 9.6 Krylov Subspace Methods

So far, we have established two broad classes of algorithms for solving linear systems. The first, known as *direct methods*, are based on some version of Gaussian Elimination or matrix factorization. Direct methods eventually<sup>†</sup> obtain the exact solution, but must be carried through to completion before any useful information is obtained. The second class contains the *iterative methods* discussed above that lead to closer and closer approximations to the solution, but almost never reach the exact value. One might ask whether there are algorithms that combine the best of both: *semi-direct methods* whose intermediate computations lead to closer and closer approximations, and, moreover, are guaranteed to terminate in a finite number of steps with the exact solution in hand.

In recent years, for dealing with large sparse linear systems, such as those arising from the numerical solution of partial differential equations, semi-direct iterative methods based on Krylov subspaces have become quite popular. The original ideas were introduced in the 1930's by the Russian naval engineer Alexei Krylov, who was in search of an efficient and reliable method for numerically computing eigenvalues. Krylov methods have seen much development in a variety of directions, [32, 70, 85], and we will show how they can be used to iteratively solve linear systems and to compute eigenvalues.

### Krylov Subspaces

The starting point is an  $n \times n$  matrix  $A$ , assumed to be real, although extensions to complex matrices are relatively straightforward. In applications,  $A$  is both large and sparse, meaning that most of its entries are 0, and so multiplying  $A$  by a vector  $\mathbf{v} \in \mathbb{R}^n$  to produce the vector  $A\mathbf{v}$  is an efficient operation.

Recall that the Power Method for computing the dominant eigenvalue and eigenvector of  $A$  is based on successive iterates applied to a randomly chosen initial vector:  $\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, A^3\mathbf{v}, \dots$ . We will employ these particular vectors to span a collection of subspaces.

**Definition 9.49.** Given an  $n \times n$  real matrix  $A$ , the *Krylov subspace* of order  $k \geq 1$  generated by a nonzero vector  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$  is the subspace  $V^{(k)} \subset \mathbb{R}^n$  spanned by the vectors  $\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{k-1}\mathbf{v}$ . We also set  $V^{(0)} = \{\mathbf{0}\}$  by convention.

<sup>†</sup> This assumes that we are dealing with a fully accurate implementation, i.e., without round-off or other numerical error. In this discussion, numerical instability will be left aside as a separate, albeit ultimately important, concern.

For example, if  $\mathbf{v}$  is an eigenvector of  $A$ , so  $A\mathbf{v} = \lambda\mathbf{v}$ , then  $V^{(2)} = V^{(1)}$  is the one-dimensional eigenspace spanned by  $\mathbf{v}$ ; conversely, if  $V^{(2)}$  is one-dimensional, then  $\mathbf{v}$  is necessarily an eigenvector, and hence  $V^{(k)} = V^{(1)}$  for all  $k \geq 1$ . More generally, if  $V^{(j+1)} = V^{(j)}$  for some  $j \geq 0$ , then  $V^{(k)} = V^{(j)}$  for all  $k \geq j$ . This is easily proved by induction: by assumption,  $A^j\mathbf{v} \in V^{(j)}$ , and thus can be written as a linear combination

$$A^j\mathbf{v} = c_1\mathbf{v} + c_2A\mathbf{v} + \cdots + c_{j-1}A^{j-2}\mathbf{v} + c_jA^{j-1}\mathbf{v} \in V^{(j)}$$

for some scalars  $c_1, \dots, c_j$ . Thus,

$$\begin{aligned} A^{j+1}\mathbf{v} &= c_1A\mathbf{v} + c_2A^2\mathbf{v} + \cdots + c_{j-1}A^{j-1}\mathbf{v} + c_jA^j\mathbf{v} \\ &= c_jc_1\mathbf{v} + (c_1 + c_jc_2)A\mathbf{v} + \cdots + (c_{j-2} + c_jc_{j-1})A^{j-2}\mathbf{v} + (c_j + c_j^2)A^{j-1}\mathbf{v} \in V^{(j)} \end{aligned}$$

also, proving that  $V^{(j+2)} = V^{(j)}$ . The general induction step is clear.

Since we assumed  $\mathbf{v} \neq \mathbf{0}$ , as otherwise all  $V^{(k)} = \{\mathbf{0}\}$  are trivial and not of interest, this argument implies the existence of an integer  $m \in \mathbb{N}$ , called the *stabilization order*, such that  $\dim V^{(k)} = k$  for  $k = 1, \dots, m$ , while  $V^{(k)} = V^{(m)}$  has dimension  $m$  for all  $k \geq m$ . Since we are working in  $\mathbb{R}^n$ , clearly  $m \leq n$ ; Exercise 9.6.3 gives a stricter bound for  $m$  in terms of the degree of the minimal polynomial of the matrix  $A$ , as defined in Exercise 8.6.23. We also note the following useful result.

**Lemma 9.50.** Suppose  $V^{(k)} \neq V^{(k-1)}$ . Let  $\mathbf{w} \in V^{(k)} \setminus V^{(k-1)}$ . Then  $A\mathbf{w} \in V^{(k+1)}$  and, moreover,  $V^{(k+1)}$  is spanned by  $A\mathbf{w}$  and (a basis of)  $V^{(k)}$ . Moreover, if  $A\mathbf{w} \in V^{(k)}$ , then  $V^{(k+1)} = V^{(k)}$  and the Krylov subspaces stabilize at order  $k$ .

*Proof:* By assumption,

$$\mathbf{w} = c_1\mathbf{v} + c_2A\mathbf{v} + \cdots + c_{k-1}A^{k-2}\mathbf{v} + c_kA^{k-1}\mathbf{v}$$

for some scalars  $c_1, \dots, c_k$  with  $c_k \neq 0$ . Thus, as above,

$$A\mathbf{w} = c_1A\mathbf{v} + c_2A^2\mathbf{v} + \cdots + c_{k-1}A^{k-1}\mathbf{v} + c_kA^k\mathbf{v} \in V^{(k+1)}. \quad (9.91)$$

If  $A\mathbf{w} \in V^{(k)}$ , the left-hand side of (9.91) is a linear combination of  $\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{k-1}\mathbf{v}$ , and hence, since  $c_k \neq 0$ , so is  $A^k\mathbf{v}$ , which implies  $V^{(k+1)} = V^{(k)}$ . Otherwise, (9.91) implies that  $A^k\mathbf{v}$  is a linear combination of  $A\mathbf{w}$  and  $A\mathbf{v}, A^2\mathbf{v}, \dots, A^{k-1}\mathbf{v}$ , and thus every vector in  $V^{(k+1)}$  can be written as a linear combination of  $A\mathbf{w}$  and the Krylov vectors  $\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \dots, A^{k-1}\mathbf{v} \in V^{(k)}$ . *Q.E.D.*

For simplicity in what follows, we will assume that  $A$  has all real eigenvalues; for example  $A$  might be a symmetric matrix. We further assume that  $A$  has a unique dominant eigenvalue  $\lambda_1$ , so that  $\lambda_1$  is a simple eigenvalue, and  $|\lambda_1| > |\lambda_j|$  for all  $j > 1$ . In this case, as we know from our earlier analysis, for most initial choices of the vector  $\mathbf{v}$ , the vectors used to define the Krylov subspace tend to scalar multiples of a dominant eigenvector  $\mathbf{v}_1$ , meaning that  $A^k\mathbf{v} \rightarrow \lambda_1^k\mathbf{v}_1$  as  $k \rightarrow \infty$ . Thus, the Krylov vectors in and of themselves contain increasingly little information, particularly in a numerical environment. As with the Power Method, matrices with several dominant eigenvalues, including real matrices with complex conjugate eigenvalues and matrices for which  $\pm\lambda_1$  are both eigenvalues, require suitable modifications of the methods.

## Arnoldi Iteration

The way to get around the pure power behavior was already introduced in the design of the  $QR$  algorithm: instead of the Krylov vectors, one constructs an orthonormal basis of

the Krylov subspace using the Gram–Schmidt process. (As above, we work with the dot product  $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \mathbf{w}$  and corresponding Euclidean norm throughout this presentation, leaving the investigation of other inner products to the motivated reader.) To this end, we may as well start with a unit vector, and so replace the initial vector  $\mathbf{v}$  by the unit vector  $\mathbf{u}_1 = \mathbf{v}/\|\mathbf{v}\|$ , so  $\|\mathbf{u}_1\| = 1$ , which spans the initial Krylov subspace  $V^{(1)}$ . The second order subspace  $V^{(2)}$  will be spanned by the vectors  $\mathbf{u}_1$  and  $A\mathbf{u}_1$ , and we extract an orthonormal basis by projection. First, according to our orthogonal projection formulas, the vector

$$\mathbf{v}_2 = A\mathbf{u}_1 - h_{11}\mathbf{u}_1, \quad \text{where} \quad h_{11} = \mathbf{u}_1^T A\mathbf{u}_1,$$

satisfies the desired orthogonality condition  $\mathbf{u}_1 \cdot \mathbf{v}_2 = 0$ . If  $\mathbf{v}_2 = \mathbf{0}$ , then  $\mathbf{u}_1$  is an eigenvector of  $A$ , and the process terminates, since the Krylov subspaces would immediately stabilize:  $V^{(k)} = V^{(1)}$  for all  $k \geq 1$ . Otherwise, we replace  $\mathbf{v}_2$  by the unit vector

$$\mathbf{u}_2 = \frac{\mathbf{v}_2}{h_{21}}, \quad \text{where} \quad h_{21} = \|\mathbf{v}_2\|,$$

and deduce that  $\mathbf{u}_1$  and  $\mathbf{u}_2$  form an orthonormal basis for  $V^{(2)}$ . Proceeding in this manner, assuming that  $k \leq m$ , the stabilization order, at the  $k^{\text{th}}$  stage, we have already computed orthonormal vectors  $\mathbf{u}_1, \dots, \mathbf{u}_k$  such that  $\mathbf{u}_1, \dots, \mathbf{u}_j$  form an orthonormal basis of  $V^{(j)}$  for each  $j = 1, \dots, k$ . Taking  $\mathbf{w} = \mathbf{u}_k$  in Lemma 9.50, we deduce that  $\mathbf{u}_1, \dots, \mathbf{u}_k$  and  $A\mathbf{u}_k$  span  $V^{(k+1)}$ . Our orthogonal projection formula (4.41) implies that

$$\mathbf{v}_{k+1} = A\mathbf{u}_k - \sum_{j=1}^k h_{jk}\mathbf{u}_j, \quad \text{where} \quad h_{jk} = \mathbf{u}_j^T A\mathbf{u}_k \quad (9.92)$$

lies in  $V^{(k+1)}$  and is orthogonal to  $\mathbf{u}_0, \dots, \mathbf{u}_k$ . If  $\mathbf{v}_{k+1} = \mathbf{0}$ , then  $A\mathbf{u}_k \in V^{(k)}$ , and, again by Lemma 9.50, the Krylov spaces have stabilized with  $V^{(k+1)} = V^{(k)}$ . Otherwise, let

$$\mathbf{u}_{k+1} = \frac{\mathbf{v}_{k+1}}{h_{k+1,k}}, \quad \text{where} \quad h_{k+1,k} = \|\mathbf{v}_{k+1}\|, \quad (9.93)$$

be the corresponding unit vector, so that  $\mathbf{u}_0, \dots, \mathbf{u}_{k+1}$  form an orthonormal basis of  $V^{(k+1)}$ , as desired.

While the preceding algorithm will work in favorable situations, the preferred method, known as *Arnoldi iteration*, named after the mid-twentieth-century American engineer Walter Arnoldi, employs the stabilized Gram–Schmidt process described in Section 4.2, thereby ameliorating, as much as possible, potential numerical instabilities. Thus, at step  $k \geq 1$ , having  $\mathbf{u}_1, \dots, \mathbf{u}_k$  in hand, one iteratively computes

$$\mathbf{v}_{k+1}^{(1)} = A\mathbf{u}_k, \quad \mathbf{v}_{k+1}^{(j+1)} = \mathbf{v}_{k+1}^{(j)} - h_{jk}\mathbf{u}_j, \quad \text{where} \quad h_{jk} = \mathbf{u}_j^T \mathbf{v}_{k+1}^{(j)}, \quad \text{for } j = 1, \dots, k-1. \quad (9.94)$$

We then set  $\mathbf{v}_{k+1} = \mathbf{v}_{k+1}^{(k)}$  and, if it is nonzero, use (9.93) to define the next orthonormal basis vector  $\mathbf{u}_{k+1}$ . In Exercise 9.6.6 you are asked to prove that the resulting *Arnoldi vectors*  $\mathbf{u}_k$  and coefficients  $h_{jk}$  are the same as in (9.92, 93) (if computed exactly).

It is instructive to formulate the Arnoldi orthonormalization process in matrix form. First note that we can rewrite (9.92–93) as

$$A\mathbf{u}_k = \sum_{j=0}^{k+1} h_{jk}\mathbf{u}_j, \quad (9.95)$$

and hence, by orthonormality

$$h_{jk} = \begin{cases} \mathbf{u}_j^T A \mathbf{u}_k, & 1 \leq j \leq k+1, \\ 0, & j \geq k+2. \end{cases} \quad (9.96)$$

Let  $Q_k = (\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_k)$  denote the  $n \times k$  matrix whose columns are the first  $k$  Arnoldi vectors. Since these are orthonormal, it follows that

$$Q_k^T Q_k = \mathbf{I}. \quad (9.97)$$

(However, keep in mind that  $Q_k$  is a rectangular matrix, and so  $Q_k Q_k^T$  is in general *not* the identity matrix.) Let

$$H_k = \begin{pmatrix} h_{11} & h_{12} & h_{13} & h_{14} & \cdots & h_{1,k-2} & h_{1,k-1} & h_{1k} \\ h_{21} & h_{22} & h_{23} & h_{24} & \cdots & h_{2,k-2} & h_{2,k-1} & h_{2k} \\ 0 & h_{32} & h_{33} & h_{34} & \cdots & h_{3,k-2} & h_{3,k-1} & h_{3k} \\ 0 & 0 & h_{43} & h_{44} & \cdots & h_{4,k-2} & h_{4,k-1} & h_{4k} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & h_{k-1,k-2} & h_{k-1,k-1} & h_{k-1,k} \\ 0 & 0 & 0 & \cdots & 0 & 0 & h_{k,k-1} & h_{kk} \end{pmatrix} \quad (9.98)$$

be the  $k \times k$  upper Hessenberg matrix formed by the coefficients  $h_{jk}$  given in (9.96), which implies that

$$H_k = Q_k^T A Q_k. \quad (9.99)$$

In particular, if  $A$  is symmetric, then so is  $H_k$ , which implies that it is also tridiagonal. In this case, the Arnoldi algorithm is known as the *symmetric Lanczos algorithm*, after the Hungarian mathematician Cornelius Lanczos.

Equation (9.99) yields an alternative interpretation of the Arnoldi iteration as a (partial) orthogonal reduction of  $A$  to Hessenberg or, in the symmetric case, tridiagonal form. The matrix  $H_k$  can be viewed as the representation of the orthogonal projection of  $A$  onto the Krylov subspace  $V^{(k)}$  in terms of the basis formed by the Arnoldi vectors  $\mathbf{u}_1, \dots, \mathbf{u}_k$ . Thus, we can identify  $H_k$  with the (projected) action of  $A$  on the subspace  $V^{(k)}$  and, as such, its dominant eigenvalues and eigenvectors, which can be computed using the  $QR$  algorithm, are expected to form good approximations to those of  $A$  itself. Since its predecessor,  $H_{k-1}$ , coincides with the upper left  $(k-1) \times (k-1)$  submatrix of  $H_k$ , the  $QR$  factorizations of the Hessenberg coefficient matrices  $H_k$  can be speeded up by an iterative procedure; see [70] for details. One can also use Householder reflections to tridiagonalize  $H_k$  before applying  $QR$ . Of course, if  $A$  is symmetric, then, as noted above,  $H_k$  is already tridiagonal and so this step is superfluous. Moreover, if the method is carried out to the stabilization order  $m$ , the resulting Krylov subspace is invariant under  $A$ , and hence the eigenvalues of  $H_m$  coincide with those of  $A$  restricted to  $V^{(m)}$ , cf. Exercise 8.4.5. In this manner, the Arnoldi/Lanczos algorithm produces a semi-direct method for approximating eigenvalues of the matrix  $A$ . Again, the Shifted Inverse Power Method of Exercise 9.5.7 can then be used to compute each corresponding eigenvector.

We further note, as a consequence of the first equation in (9.95), the following formula relating the Arnoldi matrix  $Q_k$  to its successor  $Q_{k+1}$ :

$$A Q_k = Q_{k+1} \tilde{H}_k, \quad (9.100)$$

where

$$\tilde{H}_k = \begin{pmatrix} h_{11} & h_{12} & h_{13} & h_{14} & \cdots & h_{1,k-2} & h_{1,k-1} & h_{1k} \\ h_{21} & h_{22} & h_{23} & h_{24} & \cdots & h_{2,k-2} & h_{2,k-1} & h_{2k} \\ 0 & h_{32} & h_{33} & h_{34} & \cdots & h_{3,k-2} & h_{3,k-1} & h_{3k} \\ 0 & 0 & h_{43} & h_{44} & \cdots & h_{4,k-2} & h_{4,k-1} & h_{4k} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & h_{k-1,k-2} & h_{k-1,k-1} & h_{k-1,k} \\ 0 & 0 & 0 & \cdots & 0 & 0 & h_{k,k-1} & h_{kk} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & h_{k+1,1} \end{pmatrix} \quad (9.101)$$

is the  $(k+1) \times k$  matrix formed by appending the indicated bottom row to  $H_k$ .

Finally, we note the useful formula

$$Q_k^T \mathbf{v} = \|\mathbf{v}\| \mathbf{e}_1, \quad (9.102)$$

with  $\mathbf{e}_1 = (1, 0, 0, \dots, 0)^T \in \mathbb{R}^k$  the first standard basis vector. This is a consequence of the orthonormality of the Arnoldi vectors  $\mathbf{u}_1, \dots, \mathbf{u}_k$ , which form the rows of  $Q_k^T$ , along with the fact that  $\mathbf{v} = \|\mathbf{v}\| \mathbf{u}_1$ .

**Remark.** In numerical applications, the best results are obtained by maximizing the stabilization order of the Krylov subspaces generated by the initial vector, and so a random choice of the initial vector  $\mathbf{v}$ , or, equivalently, the initial unit vector  $\mathbf{u}_1$  is preferred so as to minimize chances of low order degeneration and consequent inaccuracies. In the unlucky event that stabilization occurs prematurely, one should restart the method with a different choice of initial vector, [70].

## The Full Orthogonalization Method

Krylov subspaces can also be applied to generate powerful semi-direct iterative algorithms for solving linear systems. There are two different approaches. The first starts with the concept of a *weak* or *Galerkin formulation* of a linear system, which is the elementary observation that that the only vector that is orthogonal to every vector in an inner product space is the zero vector; see Exercise 3.1.10(a). As above, we concentrate on the case  $V = \mathbb{R}^n$  with the standard dot product. The observation means that  $\mathbf{x} \in \mathbb{R}^n$  solves the linear system  $A\mathbf{x} = \mathbf{b}$  if and only if

$$\mathbf{v}^T (A\mathbf{x} - \mathbf{b}) = 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^n. \quad (9.103)$$

Solution techniques based on this formulation were first studied in depth, in the context of the mechanics of thin elastic plates, by the Russian engineer Boris Galerkin in the first half of the twentieth century, and often bear his name.

In the case of linear algebraic systems, the Galerkin formulation per se does not add anything to what we already know. However, it becomes important for the numerical approximation of solutions by restricting (9.103) to a smaller-dimensional subspace  $V \subset \mathbb{R}^n$ . Specifically, one seeks a vector  $\mathbf{x} \in V$  such that the Galerkin formulation (9.103) holds for all  $\mathbf{v} \in V$ . In other words, the approximate solution is the vector  $\mathbf{x} \in V$  such that the residual  $\mathbf{r} = \mathbf{b} - A\mathbf{x}$  is orthogonal to the subspace  $V$ . With a suitably inspired choice of the subspace  $V$ , the Galerkin formulation may well provide a decent approximation to the actual solution.

**Remark.** One can easily adapt the Galerkin formulation to general linear systems  $L[u] = f$ , where  $L: U \rightarrow V$  is any linear operator between vector spaces. The corresponding weak formulation, as described in Exercise 7.5.9, has become an extremely important tool in the modern mathematical analysis of differential equations, which take place in infinite-dimensional function spaces. Moreover, the restriction of the weak formulation to a finite-dimensional subspace  $V \subset U$  is the basis of the powerful finite element solution method for boundary value problems; see [8, 61] for details.

**Remark.** The question of existence and uniqueness of the Galerkin approximate solution depends upon the matrix  $A$  and the choice of subspace  $V$ . Given a basis  $\mathbf{v}_1, \dots, \mathbf{v}_k$  of  $V$ , we express  $\mathbf{x} = y_1 \mathbf{v}_1 + \dots + y_k \mathbf{v}_k = S\mathbf{y}$ , where  $S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_k)$  is the  $n \times k$  matrix whose columns are the basis vectors, while  $\mathbf{y} = (y_1, y_2, \dots, y_k)^T \in \mathbb{R}^k$  contains the coordinates of  $\mathbf{x} = S\mathbf{y} \in V$  with respect to the given basis. Then the Galerkin conditions on  $V$  can be written as

$$\mathbf{v}^T (A\mathbf{x} - \mathbf{b}) = \mathbf{v}^T (AS\mathbf{y} - \mathbf{b}) = \mathbf{0} \quad \text{for all } \mathbf{v} \in V.$$

Expressing  $\mathbf{v} = S\mathbf{z}$  for  $\mathbf{z} \in \mathbb{R}^k$  in the same fashion, this becomes

$$\mathbf{z}^T S^T (AS\mathbf{y} - \mathbf{b}) = \mathbf{z}^T (S^T AS\mathbf{y} - S^T \mathbf{b}) = \mathbf{0} \quad \text{for all } \mathbf{z} \in \mathbb{R}^k,$$

which clearly holds if and only if

$$S^T AS\mathbf{y} = S^T \mathbf{b}. \quad (9.104)$$

This is a linear system of  $k$  equations in the  $k$  unknowns  $\mathbf{y} \in \mathbb{R}^k$ . Thus, a solution exists and is uniquely determined if and only if the  $k \times k$  coefficient matrix  $S^T AS$  is nonsingular, which requires, at the very least,  $\text{rank } A \geq k$ , and places additional constraints on  $S$ .

As you may suspect, in the case of a linear algebraic system, a particularly good choice of subspace for a Galerkin approximation to the solution is a Krylov subspace. The resulting solution method is known as the *Full Orthogonalization Method*, abbreviated FOM, [70]. In detail, the method proceeds as follows. Let  $V^{(k)} \subset \mathbb{R}^n$  be the order  $k$  Krylov subspace generated by the right-hand side  $\mathbf{b}$ , and thus spanned by  $\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \dots, A^{k-1}\mathbf{b}$ . The  $k^{\text{th}}$  *Krylov approximation* to the solution  $\mathbf{x}$  is the vector  $\mathbf{x}_k \in V^{(k)}$  whose residual  $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$  satisfies the Galerkin condition of being orthogonal to the subspace:

$$\mathbf{v} \cdot \mathbf{r}_k = \mathbf{v}^T (\mathbf{b} - A\mathbf{x}_k) = \mathbf{0} \quad \text{for all } \mathbf{v} \in V^{(k)}.$$

In particular, the initial approximation is taken to be  $\mathbf{x}_0 = \mathbf{0} \in V^{(1)}$ , with residual vector  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 = \mathbf{b}$ . Moreover, Lemma 9.50 implies  $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k \in V^{(k+1)}$ . Since it is orthogonal to  $V^{(k)}$ , it must be a scalar multiple of the  $(k+1)^{\text{st}}$  Arnoldi vector:

$$\mathbf{r}_k = c_{k+1} \mathbf{u}_{k+1}, \quad \text{where} \quad c_{k+1} = \|\mathbf{r}_k\|. \quad (9.105)$$

This implies that the residual vectors are also mutually orthogonal:

$$\mathbf{r}_j \cdot \mathbf{r}_k = 0, \quad j \neq k. \quad (9.106)$$

Using the orthonormal Arnoldi basis vectors  $\mathbf{u}_1, \dots, \mathbf{u}_k \in V^{(k)}$ , which form the columns of the matrix  $Q_k$ , we write  $\mathbf{x}_k = Q_k \mathbf{y}_k$ , and hence, recalling (9.99), equation (9.104) becomes

$$Q_k^T A Q_k \mathbf{y}_k = H_k \mathbf{y}_k = Q_k^T \mathbf{b} = \|\mathbf{b}\| \mathbf{e}_1, \quad (9.107)$$

where  $H_k$  is the upper Hessenberg matrix (9.99), and we use (9.102) (with  $\mathbf{b}$  replacing  $\mathbf{v}$ , as per our initial supposition) to obtain the final expression. Solving the resulting system (9.107), assuming  $H_k$  is invertible, for  $\mathbf{y}_k = \|\mathbf{b}\| H_k^{-1} \mathbf{e}_1$  produces the  $k^{\text{th}}$  order Krylov approximation to the solution

$$\mathbf{x}_k = Q_k \mathbf{y}_k = \|\mathbf{b}\| Q_k H_k^{-1} \mathbf{e}_1. \quad (9.108)$$

Of course, in applications one does not explicitly compute the inverse  $H_k^{-1}$  but rather uses, say, its  $LU$  factorization  $H_k = L_k U_k$  (assuming regularity), coupled with forward and back substitution to solve (9.107). Moreover, according to Exercise 9.5.27, the lower unitriangular factor  $L_k$  is bidiagonal, meaning that all entries not on the diagonal or subdiagonal are zero. Of course, because the upper left  $(k-1) \times (k-1)$  entries of  $H_k$  are the same as those of its predecessor, whose factorization  $H_{k-1} = L_{k-1} U_{k-1}$  can be assumed to already be known, we can quickly factorize  $H_k$ . Namely, we write

$$H_k = \begin{pmatrix} H_{k-1} & \mathbf{f}_k \\ \mathbf{g}_k^T & h_{kk} \end{pmatrix}, \quad L_k = \begin{pmatrix} L_{k-1} & \mathbf{0} \\ \mathbf{m}_k^T & 1 \end{pmatrix}, \quad U_k = \begin{pmatrix} U_{k-1} & \mathbf{z}_k \\ \mathbf{0} & u_{kk} \end{pmatrix},$$

where  $\mathbf{f}_k, \mathbf{g}_k, \mathbf{m}_k, \mathbf{z}_k \in \mathbb{R}^{k-1}$ , while  $h_{kk}, u_{kk} \in \mathbb{R}$ . Moreover, since  $H_k$  is upper Hessenberg, both  $\mathbf{g}_k = h_{k,k-1} \mathbf{e}_{k-1}$  and  $\mathbf{m}_k = l_{k,k-1} \mathbf{e}_{k-1}$  are multiples of the  $(k-1)^{\text{st}}$  basis vector  $\mathbf{e}_{k-1} \in \mathbb{R}^{k-1}$ . Multiplying out  $H_k = L_k U_k$  implies that we need only solve a single triangular linear system, via forward substitution, along with a pair of scalar linear equations, resulting in

$$L_{k-1} \mathbf{z}_k = \mathbf{f}_k, \quad l_{k,k-1} = h_{k,k-1} / u_{k-1,k-1}, \quad u_{kk} = h_{kk} - l_{k,k-1} u_{k-1,k}. \quad (9.109)$$

**Remark.** Suppose you happen to know a good initial guess  $\mathbf{x}_0$  for the solution. The convergence can then be speeded up by setting  $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_0$ , which converts the original system to  $A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ , where  $\tilde{\mathbf{b}} = \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$  is the initial residual. On applying the FOM algorithm to the modified system, the resulting  $\tilde{\mathbf{x}}_k \in \tilde{V}^{(k)}$  in the Krylov subspaces generated by  $\tilde{\mathbf{b}}$  provide the improved approximations  $\mathbf{x}_k = \tilde{\mathbf{x}}_k + \mathbf{x}_0$  to the solution  $\mathbf{x}$  to the original system.

## The Conjugate Gradient Method

The most important case of the FOM algorithm is that in which the coefficient matrix  $A$  is symmetric, and hence, as noted above,  $H_k$  is symmetric, tridiagonal, which means that the system (9.107) can be quickly solved by the tridiagonal version of Gaussian Elimination, cf. (1.69–70). In particular, if  $A > 0$  is positive definite, then so is  $H_k > 0$ , and the resulting algorithm is known as the *Conjugate Gradient Method*, often abbreviated CG, first introduced in 1952 by Hestenes and Stiefel, [39]. It is now the most widely used method for solving linear systems with positive definite coefficient matrices, e.g., those arising in the numerical solution to boundary value problems for elliptic systems of partial differential equations, [8, 61].

There is a simpler direct way to formulate the CG algorithm, which is the one that is used in practice. First, we apply Theorems 1.29 and 1.34 to refine the factorization of the tridiagonal matrix:

$$H_k = L_k D_k L_k^T, \quad (9.110)$$

where  $L_k$  is lower unitriangular and  $D_k$  is diagonal. Let  $C_k$  be the  $k \times k$  diagonal matrix with diagonal entries  $c_j = \|\mathbf{r}_{j-1}\|$  for  $j = 1, \dots, k$ , so that, according to (9.105),

$$Q_k C_k = R_k = (\mathbf{r}_0 \ \mathbf{r}_1 \ \dots \ \mathbf{r}_{k-1})$$

is the matrix of residual vectors. Define

$$W_k = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_k) = Q_k L_k^{-T} C_k = R_k V_k, \quad (9.111)$$

where the columns  $\mathbf{w}_1, \dots, \mathbf{w}_k$  of  $W_k$  are known as the *conjugate directions*, and where

$$V_k = C_k^{-1} L_k^{-T} C_k = \begin{pmatrix} 1 & s_1 & & & \\ & 1 & s_2 & & \\ & & \ddots & \ddots & \\ & & & 1 & s_{k-1} \\ & & & & 1 \end{pmatrix}$$

is upper unitriangular. Note that, for  $j \geq 1$ , the  $(j+1)^{\text{st}}$  column of the matrix equation  $R_k = W_k V_k^{-1}$  implies

$$\mathbf{r}_j = \mathbf{w}_{j+1} - s_j \mathbf{w}_j. \quad (9.112)$$

We claim that the vectors  $\mathbf{w}_1, \dots, \mathbf{w}_k$  are *conjugate*, which means that they are mutually orthogonal with respect to the inner product<sup>†</sup>  $\langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle = \mathbf{v}^T A \mathbf{w}$  induced by  $A$ , and so

$$\langle\langle \mathbf{w}_i, \mathbf{w}_j \rangle\rangle = \mathbf{w}_i^T A \mathbf{w}_j = 0, \quad i \neq j. \quad (9.113)$$

To verify (9.113), we use (9.110, 111) to compute the corresponding Gram matrix, whose entries are the inner products:

$$W_k^T A W_k = C_k L_k^{-1} Q_k^T A Q_k L_k^{-T} C_k = C_k L_k^{-1} H_k L_k^{-T} C_k = C_k D_k C_k = C_k^2 D_k,$$

the final result being a diagonal matrix. We deduce that all the off-diagonal entries of the Gram matrix  $W_k^T A W_k$  vanish, which proves (9.113).

Let us write the  $k^{\text{th}}$  approximate solution  $\mathbf{x}_k \in V^{(k)}$  in the form

$$\mathbf{x}_k = Q_k \mathbf{y}_k = W_k \mathbf{t}_k = t_1 \mathbf{w}_1 + \dots + t_k \mathbf{w}_k, \quad \text{where} \quad \mathbf{t}_k = C_k^{-1} L_k^T \mathbf{y}_k.$$

As a consequence of (9.112) with<sup>‡</sup>  $k$  replacing  $j$ , along with (9.113), its residual vector  $\mathbf{r}_k = \mathbf{b} - A \mathbf{x}_k$  satisfies

$$\begin{aligned} \langle\langle \mathbf{r}_k, \mathbf{w}_k \rangle\rangle &= \langle\langle \mathbf{w}_{k+1} - s_k \mathbf{w}_k, \mathbf{w}_k \rangle\rangle = -s_k \langle\langle \mathbf{w}_k, \mathbf{w}_k \rangle\rangle, \\ \langle\langle \mathbf{r}_k, \mathbf{w}_{k+1} \rangle\rangle &= \langle\langle \mathbf{w}_{k+1} - s_k \mathbf{w}_k, \mathbf{w}_{k+1} \rangle\rangle = \langle\langle \mathbf{w}_{k+1}, \mathbf{w}_{k+1} \rangle\rangle. \end{aligned} \quad (9.114)$$

The  $(k+1)^{\text{st}}$  approximation can be written in the iterative form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_{k+1} \mathbf{w}_{k+1}, \quad (9.115)$$

meaning that we move from  $\mathbf{x}_k$  to  $\mathbf{x}_{k+1}$  by adding a suitable scalar multiple of the conjugate direction  $\mathbf{w}_{k+1}$ . The updated residual is

$$\mathbf{r}_{k+1} = \mathbf{b} - A \mathbf{x}_{k+1} = \mathbf{b} - A \mathbf{x}_k - t_{k+1} A \mathbf{w}_{k+1} = \mathbf{r}_k - t_{k+1} A \mathbf{w}_{k+1}. \quad (9.116)$$

---

<sup>†</sup> Of course, (9.113) defines a genuine inner product only if  $A > 0$ . On the other hand, the ensuing calculations only require symmetry of the coefficient matrix, although there is no guarantee that the resulting linear systems can be solved when  $A$  is not positive definite.

<sup>‡</sup> To be completely accurate, the resulting equation appears as the  $(k+1)^{\text{st}}$  column of the subsequent matrix equations  $R_l = W_l V_l^{-1}$  for all  $l \geq k+1$ .

---

 Conjugate Gradient Method for Solving  $A\mathbf{x} = \mathbf{b}$  with  $A > 0$ 


---



---

```

start
  choose an initial guess  $\mathbf{x}_0$ , e.g.,  $\mathbf{x}_0 = \mathbf{0}$ 
  for  $k = 0$  to  $m - 1$ 
    set  $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ 
    if  $\mathbf{r}_k = \mathbf{0}$  print “ $\mathbf{x}_k$  is the exact solution”; end
    if  $k = 0$  set  $\mathbf{w}_1 = \mathbf{r}_0$  else set  $\mathbf{w}_{k+1} = \mathbf{r}_k + \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2} \mathbf{w}_k$ 
    set  $\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\|\mathbf{r}_k\|^2}{\mathbf{w}_{k+1}^T A \mathbf{w}_{k+1}} \mathbf{w}_{k+1}$ 
    next  $k$ 
  end

```

---

Orthogonality of the residuals, (9.106), coupled with (9.114) implies

$$\begin{aligned} 0 = \mathbf{r}_k^T \mathbf{r}_{k+1} &= \|\mathbf{r}_k\|^2 - t_{k+1} \mathbf{r}_k^T A \mathbf{w}_{k+1} = \|\mathbf{r}_k\|^2 - t_{k+1} \langle\langle \mathbf{r}_k, \mathbf{w}_{k+1} \rangle\rangle \\ &= \|\mathbf{r}_k\|^2 - t_{k+1} \langle\langle \mathbf{w}_{k+1}, \mathbf{w}_{k+1} \rangle\rangle, \end{aligned}$$

hence

$$t_{k+1} = \frac{\|\mathbf{r}_k\|^2}{\langle\langle \mathbf{w}_{k+1}, \mathbf{w}_{k+1} \rangle\rangle} = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{w}_{k+1}^T A \mathbf{w}_{k+1}}. \quad (9.117)$$

Finally, using (9.106) and (9.116, 117), with  $k$  replaced by  $k - 1$ , yields

$$\|\mathbf{r}_k\|^2 = \mathbf{r}_k^T (\mathbf{r}_{k-1} - t_k A \mathbf{w}_k) = -t_k \mathbf{r}_k^T A \mathbf{w}_k = -t_k \langle\langle \mathbf{r}_k, \mathbf{w}_k \rangle\rangle = -\frac{\langle\langle \mathbf{r}_k, \mathbf{w}_k \rangle\rangle \|\mathbf{r}_{k-1}\|^2}{\langle\langle \mathbf{w}_k, \mathbf{w}_k \rangle\rangle}.$$

Thus, referring back to (9.114),

$$s_k = -\frac{\langle\langle \mathbf{r}_k, \mathbf{w}_k \rangle\rangle}{\langle\langle \mathbf{w}_k, \mathbf{w}_k \rangle\rangle} = \frac{\|\mathbf{r}_k\|^2}{\|\mathbf{r}_{k-1}\|^2}. \quad (9.118)$$

The iterative equations (9.115, 117, 118) constitute the Conjugate Gradient algorithm, which is summarized in the accompanying pseudocode. The algorithm can also be applied if  $A$  is merely symmetric, although it may break down if the denominator  $\mathbf{w}_{k+1}^T A \mathbf{w}_{k+1} = 0$ , which will not occur in the positive definite case (why?). At each stage,  $\mathbf{x}_k$  is the current approximation to the solution. The initial guess  $\mathbf{x}_0$  can be chosen by the user, with  $\mathbf{x}_0 = \mathbf{0}$  the default. The number of iterations  $m \leq n$  can be specified in advance; alternatively, one can impose a stopping criterion based on the size of the residual vector,  $\|\mathbf{r}_k\|$ , or, alternatively, the amount of change between successive iterates, as measured by, say, their distance  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$  in either the Euclidean norm or the  $\infty$  norm. Because the process is semi-direct, eventually  $\mathbf{r}_k = \mathbf{0}$  for some  $k \leq n$ , and so, in the absence of round-off errors, the result will be the exact solution to the system. Of course, in examples, one would not carry through the algorithm to the bitter end, since a decent approximation to the solution is typically obtained with relatively few iterations. For further developments and applications, see [21, 66, 70, 89].

**Remark.** The reason for the name “conjugate gradient” is as follows. The term gradient stems from the minimization principle characterizing the solutions to linear systems with

positive definite coefficient matrices. According to Theorem 5.2, if  $A > 0$ , the solution to the linear system  $A\mathbf{x} = \mathbf{b}$  is the unique minimizer of the quadratic function

$$p(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T A \mathbf{x} - \mathbf{x}^T \mathbf{b}. \quad (9.119)$$

One approach to solving the system is to try to successively minimize  $p(\mathbf{x})$  as much as possible. Suppose we find ourselves at a point  $\mathbf{x}$  that is not the minimizer. In which direction should we travel? Multivariable calculus tells us that the gradient vector  $\nabla p(\mathbf{x})$  of a function points in the direction of steepest increase at the point, while its negative  $-\nabla p(\mathbf{x})$  points in the direction of steepest decrease, [2, 78]. The gradient of the particular quadratic function (9.119) is easily found:

$$-\nabla p(\mathbf{x}) = \mathbf{b} - A\mathbf{x} = \mathbf{r}.$$

Thus, the *residual vector* specifies the direction of steepest decrease in the quadratic function, and is thus a good choice of direction in which to head off in search of the true minimizer. (If one views the graph of  $p$  as a mountain range, then, at any given location  $\mathbf{x}$  with elevation  $p(\mathbf{x})$ , the negative gradient  $-\nabla p(\mathbf{x}) = \mathbf{r}$  points in the steepest downhill direction.) This idea leads to the *gradient descent algorithm*, in which each successive approximation  $\mathbf{x}_k$  to the solution is obtained by going a certain distance in the residual direction:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + d_k \mathbf{r}_k, \quad \text{where} \quad \mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k. \quad (9.120)$$

The scalar factor  $d_k$  is to be specified so that the resulting  $p(\mathbf{x}_{k+1})$  is as small as possible; in Exercise 9.6.14 you are asked to find this value. Gradient descent is a reasonable algorithm, and will lead to the solution in favorable situations. It is also effectively used to find minima of more general nonlinear functions. However, in certain circumstances, the iterative method based on gradient descent can take a long time to converge to an accurate approximation to the solution, and so is typically not competitive. To obtain the speedier Conjugate Gradient algorithm, we modify the gradient descent idea by requiring that the next descent direction be chosen so that it is *conjugate* to the preceding directions, i.e., satisfies (9.113). This idea can be used to produce an independent direct derivation of the Conjugate Gradient algorithm.

**Example 9.51.** Consider the linear system  $A\mathbf{x} = \mathbf{b}$  with

$$A = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}.$$

The exact solution is  $\mathbf{x}_* = (2, 5, -6)^T$ . Let us implement the method of conjugate gradients, starting with the initial guess  $\mathbf{x}_0 = (0, 0, 0)^T$ . The corresponding residual vector is merely  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0 = \mathbf{b} = (1, 2, -1)^T$ . The first conjugate direction is  $\mathbf{w}_1 = \mathbf{r}_0 = (1, 2, -1)^T$ , and we use formula (9.115) to obtain the updated approximation to the solution

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{\|\mathbf{r}_0\|^2}{\langle\langle \mathbf{w}_1, \mathbf{w}_1 \rangle\rangle} \mathbf{w}_1 = \frac{6}{4} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{pmatrix},$$

noting that  $\langle\langle \mathbf{w}_1, \mathbf{w}_1 \rangle\rangle = \mathbf{w}_1^T A \mathbf{w}_1 = 4$ . For the next stage of the algorithm, we compute

the corresponding residual  $\mathbf{r}_1 = \mathbf{b} - A\mathbf{x}_1 = \left( -\frac{1}{2}, -1, -\frac{5}{2} \right)^T$ . The conjugate direction is

$$\mathbf{w}_2 = \mathbf{r}_1 + \frac{\|\mathbf{r}_1\|^2}{\|\mathbf{r}_0\|^2} \mathbf{w}_1 = \begin{pmatrix} -\frac{1}{2} \\ -1 \\ -\frac{5}{2} \end{pmatrix} + \frac{15}{6} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix},$$

which, as designed, satisfies the conjugacy condition  $\langle\langle \mathbf{w}_1, \mathbf{w}_2 \rangle\rangle = \mathbf{w}_1^T A \mathbf{w}_2 = 0$ . Each entry of the ensuing approximation

$$\mathbf{x}_2 = \mathbf{x}_1 + \frac{\|\mathbf{r}_1\|^2}{\langle\langle \mathbf{w}_2, \mathbf{w}_2 \rangle\rangle} \mathbf{w}_2 = \begin{pmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{pmatrix} + \frac{15}{\frac{27}{4}} \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix} = \begin{pmatrix} \frac{7}{3} \\ \frac{14}{3} \\ -\frac{17}{3} \end{pmatrix} \simeq \begin{pmatrix} 2.3333 \\ 4.6667 \\ -5.6667 \end{pmatrix}$$

is now within a  $\frac{1}{3}$  of the exact solution  $\mathbf{x}_*$ .

Since we are dealing with a  $3 \times 3$  system, we will recover the exact solution by one more iteration of the algorithm. The new residual is  $\mathbf{r}_2 = \mathbf{b} - A\mathbf{x}_2 = \left( -\frac{4}{3}, \frac{2}{3}, 0 \right)^T$ . The final conjugate direction is

$$\mathbf{w}_3 = \mathbf{r}_2 + \frac{\|\mathbf{r}_2\|^2}{\|\mathbf{r}_1\|^2} \mathbf{w}_2 = \begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \end{pmatrix} + \frac{20}{\frac{15}{2}} \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix} = \begin{pmatrix} -\frac{10}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix},$$

which, as you can check, is conjugate to both  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . The solution is obtained from

$$\mathbf{x}_3 = \mathbf{x}_2 + \frac{\|\mathbf{r}_2\|^2}{\langle\langle \mathbf{w}_3, \mathbf{w}_3 \rangle\rangle} \mathbf{w}_3 = \begin{pmatrix} \frac{7}{3} \\ \frac{14}{3} \\ -\frac{17}{3} \end{pmatrix} + \frac{20}{\frac{200}{27}} \begin{pmatrix} -\frac{10}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ -6 \end{pmatrix}.$$

## The Generalized Minimal Residual Method

A natural alternative to the Galerkin weak approach is to try to directly minimize the norm of the residual  $\mathbf{r} = \mathbf{b} - A\mathbf{x}$  when the approximate solution  $\mathbf{x}$  is required to lie in a specified subspace  $\mathbf{x} \in V$ . When  $V$  is a Krylov subspace, this idea results in the Generalized Minimal Residual Method (usually abbreviated GMRES), which was developed by the Algerian and American mathematicians/computer scientists<sup>†</sup> Yousef Saad and Martin Schultz, [71].

As in the FOR Method, we choose the Krylov subspaces generated by  $\mathbf{b}$ , the right-hand side of the system to be solved, but now seek the vector  $\mathbf{x}_k^* \in V^{(k)}$  that minimizes the Euclidean norm  $\|A\mathbf{x} - \mathbf{b}\|$  over all vectors  $\mathbf{x} \in V^{(k)}$ . This approach corresponds to the initial approximation  $\mathbf{x}_0 = \mathbf{0} \in V^{(1)}$ ; as before, if we know a better initial guess  $\mathbf{x}_0$ , we set  $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_0$ , which converts the original system to  $A\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ , where  $\tilde{\mathbf{b}} = \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$  is the initial residual, and then apply the method to the new system.

Again, we express the vectors

$$\mathbf{x}_k = y_1 \mathbf{u}_1 + \cdots + y_k \mathbf{u}_k = Q_k \mathbf{y} \in V^{(k)}$$

---

<sup>†</sup> Coincidentally, the first author of the book you are reading is at the same university, Minnesota, as Saad, and had the same thesis advisor, Garrett Birkhoff, as Schultz.

as linear combinations of the orthonormal Arnoldi basis vectors, with coefficients  $\mathbf{y}_k = (y_1, \dots, y_k)^T \in \mathbb{R}^k$ . In view of (9.100) and (9.97), with  $k$  replaced by  $k+1$ , the squared residual norm is given by

$$\begin{aligned}\|\mathbf{r}\|_k^2 &= \|A\mathbf{x}_k - \mathbf{b}\|^2 = \|AQ_k\mathbf{y}_k - \mathbf{b}\|^2 = \|Q_{k+1}\tilde{H}_k\mathbf{y}_k - \mathbf{b}\|^2 \\ &= (Q_{k+1}\tilde{H}_k\mathbf{y}_k - \mathbf{b})^T(Q_{k+1}\tilde{H}_k\mathbf{y}_k - \mathbf{b}) = \mathbf{y}_k^T\tilde{H}_k^T\tilde{H}_k\mathbf{y}_k - 2\mathbf{y}_k^T\tilde{H}_k^TQ_{k+1}^T\mathbf{b} + \|\mathbf{b}\|^2 \\ &= \mathbf{y}_k^T\tilde{H}_k^T\tilde{H}_k\mathbf{y}_k - 2\mathbf{y}_k^T\tilde{H}_k^T\mathbf{c}_k + \|\mathbf{c}_k\|^2 = \|\tilde{H}_k\mathbf{y}_k - \mathbf{c}_k\|^2,\end{aligned}\quad (9.121)$$

where, according to (9.107) again with  $k$  replaced by  $k+1$ ,

$$\mathbf{c}_k = Q_{k+1}^T\mathbf{b} = \|\mathbf{b}\|\mathbf{e}_1 \in \mathbb{R}^{k+1}, \quad \text{so that} \quad \|\mathbf{c}_k\| = \|\mathbf{b}\|. \quad (9.122)$$

We deduce that minimizing  $\|A\mathbf{x} - \mathbf{b}\|$  over all  $\mathbf{x} \in V^{(k)}$  is the same as minimizing  $\|\tilde{H}_k\mathbf{y} - \mathbf{c}_k\|$  over all  $\mathbf{y} \in \mathbb{R}^k$ . The latter is a standard least squares minimization problem, whose solution  $\mathbf{y}_k$  is found by solving the corresponding normal equations

$$\tilde{H}_k^T\tilde{H}_k\mathbf{y}_k = \tilde{H}_k^T\mathbf{c}_k = \|\mathbf{b}\|\tilde{H}_k^T\mathbf{e}_1 = \|\mathbf{b}\|(h_{11}, h_{12}, \dots, h_{1k})^T. \quad (9.123)$$

Solving (9.123), produces the desired minimizer  $\mathbf{x}_k = Q_k\mathbf{y}_k \in V^{(k)}$ , and hence the desired approximation to the solution to the original linear system.

The result of this calculation is the *Generalized Minimal Residual Method* (GMRES) algorithm. To successively approximate the solution to  $A\mathbf{x} = \mathbf{b}$ , on the  $k^{\text{th}}$  iteration, we set  $\mathbf{c} = \|\mathbf{b}\|\mathbf{e}_1$ , and then perform the following steps:

- (a) calculate  $\mathbf{u}_k$  and  $\tilde{H}_k$  using the Arnoldi Method;
- (b) use least squares to find the vector  $\mathbf{y} = \mathbf{y}_k$  that minimizes  $\|\tilde{H}_k\mathbf{y} - \mathbf{c}\|$ ;
- (c) let  $\mathbf{x}_k = Q_k\mathbf{y}_k$  be the  $k^{\text{th}}$  approximate solution.

The process is repeated until the residual norm  $\|\mathbf{r}_k\| = \|A\mathbf{x}_k - \mathbf{b}\| = \|\tilde{H}_k\mathbf{y} - \mathbf{c}\|$  is below a pre-assigned threshold. Again, because of the iterative structure of the Krylov vectors, and hence the upper Hessenberg matrices  $H_k$ , knowing the solution to the order  $k$  minimization problem allows one to rather quickly construct that of the order  $k+1$  version. As with all Krylov methods, GMRES is a semi-direct method and hence, if performed in exact arithmetic, will eventually produce the exact solution once the Krylov stabilization order is reached. As with FOM/CG, this is rarely required, and one typically imposes a stopping criterion based on either the norm of the residual vector or the size of the difference between successive iterates  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ . The method works very well in practice, particularly with the sparse coefficient matrices arising in many numerical solution algorithms for partial differential equations and beyond, including finite difference, finite element, collocation, and multipole expansion.

## Exercises

9.6.1. Find an orthonormal basis for the Krylov subspaces  $V^{(1)}, V^{(2)}, V^{(3)}$  for the following matrices and vectors:

$$(a) A = \begin{pmatrix} 0 & 1 \\ 3 & 1 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}; \quad (b) A = \begin{pmatrix} 2 & 2 & -1 \\ 2 & -1 & 0 \\ 2 & 1 & 3 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix};$$

$$(c) A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 2 & -3 \\ 2 & -1 & 0 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}; \quad (d) A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

9.6.2. Let  $\mathbf{v} = \mathbf{x} + i\mathbf{y}$  be an eigenvector corresponding to a complex, non-real eigenvalue of the real  $n \times n$  matrix  $A$ . (a) Prove that the Krylov subspaces  $V^{(k)}$  for  $k \geq 2$  generated by both  $\mathbf{x}$  and  $\mathbf{y}$  are all two-dimensional. (b) Is the converse valid? Specifically, if  $\dim V^{(3)} = 2$ , then all  $V^{(k)}$  are two-dimensional for  $k \geq 1$  and spanned by the real and imaginary parts of a complex eigenvector of  $A$ .

◇ 9.6.3. (a) Prove that the dimension of a Krylov subspace is bounded by the degree of the minimal polynomial of the matrix  $A$ , as defined in Exercise 8.6.23. (b) Is there always a Krylov subspace whose dimension equals the degree of the minimal polynomial?

9.6.4. *True or false:* A Krylov subspace is an invariant subspace for the matrix  $A$ .

9.6.5. Prove that the invertibility of the coefficient matrix  $S^T A S$  in (9.104) depends only on the subspace  $V$  and not on the choice of basis thereof.

◇ 9.6.6. Prove that (9.92, 93, 94) give the same Arnoldi vectors  $\mathbf{u}_k$  and the same coefficients  $h_{jk}$  when computed exactly.

9.6.7. Solve the following linear systems by the Conjugate Gradient Method, keeping track of the residual vectors and solution approximations as you iterate.

$$(a) \begin{pmatrix} 3 & -1 \\ -1 & 5 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad (b) \begin{pmatrix} 6 & 2 & 1 \\ 2 & 3 & -1 \\ 1 & -1 & 2 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}, \quad (c) \begin{pmatrix} 6 & -1 & -3 \\ -1 & 7 & 4 \\ -3 & 4 & 9 \end{pmatrix} \mathbf{u} = \begin{pmatrix} -1 \\ -2 \\ 7 \end{pmatrix},$$

$$(d) \begin{pmatrix} 6 & -1 & -1 & 5 \\ -1 & 7 & 1 & -1 \\ -1 & 1 & 3 & -3 \\ 5 & -1 & -3 & 6 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}, \quad (e) \begin{pmatrix} 5 & 1 & 1 & 1 \\ 1 & 5 & 1 & 1 \\ 1 & 1 & 5 & 1 \\ 1 & 1 & 1 & 5 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 4 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

♣ 9.6.8. Use the Conjugate Gradient Method to solve the system in Exercise 9.4.33. How many iterations do you need to obtain the solution that is accurate to 2 decimal places? How does this compare to the Jacobi and SOR Methods?

♣ 9.6.9. According to Example 3.39, the  $n \times n$  Hilbert matrix  $H_n$  is positive definite, and hence we can apply the Conjugate Gradient Method to solve the linear system  $H_n \mathbf{u} = \mathbf{f}$ . For the values  $n = 5, 10, 30$ , let  $\mathbf{u}^* \in \mathbb{R}^n$  be the vector with all entries equal to 1.

- (a) Compute  $\mathbf{f} = H_n \mathbf{u}^*$ . (b) Use Gaussian Elimination to solve  $H_n \mathbf{u} = \mathbf{f}$ . How close is your solution to  $\mathbf{u}^*$ ? (c) Does pivoting improve the solution in part (b)?  
(d) Does the conjugate gradient algorithm do any better?

9.6.10. Try applying the Conjugate Gradient algorithm to the system  $-x + 2y + z = -2$ ,  $y + 2z = 1$ ,  $3x + y - z = 1$ . Do you obtain the solution? Why or why not?

9.6.11. *True or false:* If the residual vector  $\mathbf{r} = \mathbf{b} - A\mathbf{x}$  satisfies  $\|\mathbf{r}\| < .01$ , then  $\mathbf{x}$  approximates the true solution to within two decimal places.

◇ 9.6.12. How many arithmetic operations are needed to implement one iteration of the Conjugate Gradient Method? How many iterations can you perform before the method becomes more work than direct Gaussian Elimination?

**Remark.** If the matrix is sparse, the number of operations can decrease dramatically.

◇ 9.6.13. Fill in the details in a direct derivation of the Conjugate Gradient algorithm following the ideas outlined in the text: starting with the initial guess  $\mathbf{x}_0$  and corresponding residual vector  $\mathbf{w}_1 = \mathbf{r}_0 = \mathbf{b}$ , at the  $k^{\text{th}}$  step in the algorithm, given the approximation  $\mathbf{x}_k$  and residual  $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ , the  $k^{\text{th}}$  conjugate direction is chosen so that  $\mathbf{w}_{k+1} = \mathbf{r}_k + s_k \mathbf{w}_k$  satisfies the conjugacy conditions (9.113). The next approximation  $\mathbf{x}_{k+1} = \mathbf{x}_k + t_{k+1} \mathbf{w}_{k+1}$  is chosen so that its residual  $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1}$  is as small as possible.

◇ 9.6.14. In (9.120), find the value of  $d_k$  that minimizes  $p(\mathbf{x}_{k+1})$ .

- ♣ 9.6.15. Use the direct gradient descent algorithm (9.120) using the value of  $d_k$  found in Exercise 9.6.14 to solve the linear systems in Exercise 9.6.7. Compare the speed of convergence with that of the Conjugate Gradient Method.
- ♣ 9.6.16. Use GMRES to solve the system in Exercise 9.4.33. Compare the rate of convergence with the CG algorithm in Exercise 9.6.8.
- ♣ 9.6.17. Is GMRES able to solve the system in Exercise 9.6.10?
- 9.6.18. Explain in what sense the GMRES approximation  $\mathbf{x}_{k+1}$  of order  $k+1$  is a better approximation to the true solution than that of order  $k$ , namely  $\mathbf{x}_k$ .
- 9.6.19. (a) Explain what happens to the GMRES algorithm if the right-hand side  $\mathbf{b}$  of the linear system  $A\mathbf{x} = \mathbf{b}$  is an eigenvector of  $A$ . (b) More generally, prove that if the Krylov subspaces generated by  $\mathbf{b}$  stabilize at order  $m$ , then the solution of the linear system lies in  $V^{(m)}$  and so the GMRES algorithm converges to the solution at order  $m$ .

## 9.7 Wavelets

Trigonometric Fourier series, both continuous and discrete, are amazingly powerful, but they do suffer from one potentially serious defect. The complex exponential basis functions  $e^{ikx} = \cos kx + i \sin kx$  are spread out over the entire interval  $[-\pi, \pi]$ , and so are not well suited to processing localized signals — meaning data that are concentrated in a relatively small regions. Ideally, one would like to construct a system of functions that is orthogonal, and so has all the advantages of the Fourier basis functions, but, in addition, adapts to localized structures in signals. This dream was the inspiration for the development of the modern theory of wavelets.

### The Haar Wavelets

Although the modern era of wavelets started in the mid 1980's, the simplest example of a wavelet basis was discovered by the Hungarian mathematician Alfréd Haar in 1910, [35]. We consider the space of functions (signals) defined the interval  $[0, 1]$ , equipped with the standard  $L^2$  inner product

$$\langle f, g \rangle = \int_0^1 f(x) g(x) dx. \quad (9.124)$$

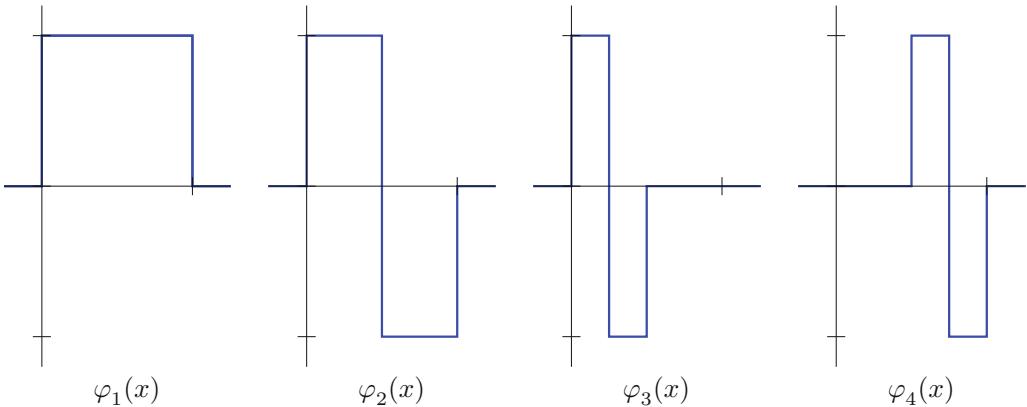
The usual scaling arguments can be used to adapt the wavelet formulas to any other interval.

The *Haar wavelets* are certain piecewise constant functions. The initial four are graphed in [Figure 9.6](#). The first is the *box function*

$$\varphi_1(x) = \varphi(x) = \begin{cases} 1, & 0 < x \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (9.125)$$

known as the *scaling function*, for reasons that shall appear shortly. Although we are interested in the value of  $\varphi(x)$  only on the interval  $[0, 1]$ , it will be convenient to extend it, and all the other wavelets, to be zero outside the basic interval. The second Haar function

$$\varphi_2(x) = w(x) = \begin{cases} 1, & 0 < x \leq \frac{1}{2}, \\ -1, & \frac{1}{2} < x \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (9.126)$$



**Figure 9.6.** The First Four Haar Wavelets.

is known as the *mother wavelet*. The third and fourth Haar functions are compressed versions of the mother wavelet:

$$\varphi_3(x) = w(2x) = \begin{cases} 1, & 0 < x \leq \frac{1}{4}, \\ -1, & \frac{1}{4} < x \leq \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases} \quad \varphi_4(x) = w(2x - 1) = \begin{cases} 1, & \frac{1}{2} < x \leq \frac{3}{4}, \\ -1, & \frac{3}{4} < x \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

called *daughter wavelets*. One can easily check, by direct evaluation of the integrals, that the four Haar wavelet functions are orthogonal with respect to the  $L^2$  inner product (9.124):  $\langle \varphi_i, \varphi_j \rangle = 0$  when  $i \neq j$ .

The scaling transformation  $x \mapsto 2x$  serves to compress the wavelet function, while the translation  $2x \mapsto 2x - 1$  moves the compressed version to the right by a half a unit. Furthermore, we can represent the mother wavelet by compressing and translating the scaling function:

$$w(x) = \varphi(2x) - \varphi(2x - 1). \quad (9.127)$$

It is these two operations of scaling and compression — coupled with the all-important orthogonality — that underlies the power of wavelets.

The Haar wavelets have an evident discretization. If we decompose the interval  $(0, 1]$  into the four subintervals

$$(0, \frac{1}{4}], \quad (\frac{1}{4}, \frac{1}{2}], \quad (\frac{1}{2}, \frac{3}{4}], \quad (\frac{3}{4}, 1], \quad (9.128)$$

on which the four wavelet functions are constant, then we can represent each of them by a vector in  $\mathbb{R}^4$  whose entries are the values of each wavelet function sampled at the left endpoint of each subinterval. In this manner, we obtain the wavelet sample vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}. \quad (9.129)$$

which form the orthogonal wavelet basis of  $\mathbb{R}^4$  we first encountered in Examples 2.35 and 4.10. Orthogonality of the vectors (9.129) with respect to the standard Euclidean dot product is equivalent to orthogonality of the Haar wavelet functions with respect to the

inner product (9.124). Indeed, if

$$f(x) \sim \mathbf{f} = (f_1, f_2, f_3, f_4) \quad \text{and} \quad g(x) \sim \mathbf{g} = (g_1, g_2, g_3, g_4)$$

are *piecewise constant* real functions that achieve the indicated values on the four subintervals (9.128), then their  $L^2$  inner product

$$\langle f, g \rangle = \int_0^1 f(x) g(x) dx = \frac{1}{4} (f_1 g_1 + f_2 g_2 + f_3 g_3 + f_4 g_4) = \frac{1}{4} \mathbf{f} \cdot \mathbf{g},$$

is equal to the averaged dot product of their sample values — the real form of the inner product (5.104) that was used in the discrete Fourier transform.

Since the vectors (9.129) form an orthogonal basis of  $\mathbb{R}^4$ , we can uniquely decompose such a piecewise constant function as a linear combination of wavelets

$$f(x) = c_1 \varphi_1(x) + c_2 \varphi_2(x) + c_3 \varphi_3(x) + c_4 \varphi_4(x),$$

or, equivalently, in terms of the sample vectors,

$$\mathbf{f} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 + c_4 \mathbf{v}_4.$$

The required coefficients

$$c_k = \frac{\langle f, \varphi_k \rangle}{\|\varphi_k\|^2} = \frac{\mathbf{f} \cdot \mathbf{v}_k}{\|\mathbf{v}_k\|^2}$$

are fixed by our usual orthogonality formula (4.7). Explicitly,

$$\begin{aligned} c_1 &= \frac{1}{4} (f_1 + f_2 + f_3 + f_4), & c_3 &= \frac{1}{2} (f_1 - f_2), \\ c_2 &= \frac{1}{4} (f_1 + f_2 - f_3 - f_4), & c_4 &= \frac{1}{2} (f_3 - f_4). \end{aligned}$$

Before proceeding to the more general case, let us introduce an important analytical definition that quantifies precisely how localized a function is.

**Definition 9.52.** The *support* of a function  $f(x)$ , written  $\text{supp } f$ , is the closure of the set where  $f(x) \neq 0$ .

Thus, a point will belong to the support of  $f(x)$ , if  $f$  is not zero there, or at least is not zero at nearby points. More precisely:

**Lemma 9.53.** If  $f(a) \neq 0$ , then  $a \in \text{supp } f$ . More generally, a point  $a \in \text{supp } f$  if and only if there exists a convergent sequence  $x_n \rightarrow a$  such that  $f(x_n) \neq 0$ . Conversely,  $a \notin \text{supp } f$  if and only if  $f(x) \equiv 0$  on an interval  $a - \delta < x < a + \delta$  for some  $\delta > 0$ .

Intuitively, the smaller the support of a function, the more localized it is. For example, the support of the Haar mother wavelet (9.126) is  $\text{supp } w = [0, 1]$  — the point  $x = 0$  is included, even though  $w(0) = 0$ , because  $w(x) \neq 0$  at nearby points. The two daughter wavelets have smaller support:

$$\text{supp } \varphi_3 = \left[0, \frac{1}{2}\right], \quad \text{supp } \varphi_4 = \left[\frac{1}{2}, 1\right],$$

and so are twice as localized.

The effect of scalings and translations on the support of a function is easily discerned.

**Lemma 9.54.** If  $\text{supp } f = [a, b]$ , and

$$g(x) = f(rx - \delta), \quad \text{then} \quad \text{supp } g = \left[\frac{a + \delta}{r}, \frac{b + \delta}{r}\right].$$

In other words, scaling  $x$  by a factor  $r$  compresses the support of the function by a factor  $1/r$ , while translating  $x$  translates the support of the function.

The key requirement for a wavelet basis is that it contains functions with arbitrarily small support. To this end, the full Haar wavelet basis is obtained from the mother wavelet by iterating the scaling and translation processes. We begin with the scaling function

$$\varphi(x), \quad (9.130)$$

from which we construct the mother wavelet via (9.127). For each “generation”  $j \geq 0$ , we form the wavelet offspring by first compressing the mother wavelet so that its support fits into an interval of length  $2^{-j}$ ,

$$w_{j,0}(x) = w(2^j x), \quad \text{so that} \quad \text{supp } w_{j,0} = [0, 2^{-j}], \quad (9.131)$$

and then translating  $w_{j,0}$  so as to fill up the entire interval  $[0, 1]$  by  $2^j$  subintervals, each of length  $2^{-j}$ , defining

$$w_{j,k}(x) = w_{j,0}(x - k) = w(2^j x - k), \quad \text{where} \quad k = 0, 1, \dots, 2^j - 1. \quad (9.132)$$

Lemma 9.54 implies that  $\text{supp } w_{j,k} = [2^{-j}k, 2^{-j}(k+1)]$ , and so the combined supports of all the  $j^{\text{th}}$  generation of wavelets is the entire interval:  $\bigcup_{k=0}^{2^j-1} \text{supp } w_{j,k} = [0, 1]$ . The primal generation,  $j = 0$ , consists of just the mother wavelet

$$w_{0,0}(x) = w(x).$$

The first generation,  $j = 1$ , consists of the two daughter wavelets already introduced as  $\varphi_3$  and  $\varphi_4$ , namely

$$w_{1,0}(x) = w(2x), \quad w_{1,1}(x) = w(2x - 1).$$

The second generation,  $j = 2$ , appends four additional granddaughter wavelets to our basis:

$$w_{2,0}(x) = w(4x), \quad w_{2,1}(x) = w(4x - 1), \quad w_{2,2}(x) = w(4x - 2), \quad w_{2,3}(x) = w(4x - 3).$$

The 8 Haar wavelets  $\varphi, w_{0,0}, w_{1,0}, w_{1,1}, w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}$  are constant on the 8 subintervals of length  $\frac{1}{8}$ , taking the successive sample values indicated by the columns of the *wavelet matrix*

$$W_8 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}. \quad (9.133)$$

Orthogonality of the wavelets is manifested in the orthogonality of the columns of  $W_8$ . (Unfortunately, terminological constraints prevent us from calling  $W_8$  an orthogonal matrix, because its columns are not orthonormal!)

The  $n^{\text{th}}$  stage consists of  $2^{n+1}$  different wavelet functions comprising the scaling functions and all the generations up to the  $n^{\text{th}}$ :  $w_0(x) = \varphi(x)$  and  $w_{j,k}(x)$  for  $0 \leq j \leq n$  and  $0 \leq k < 2^j$ . They are all constant on each subinterval of length  $2^{-n-1}$ .

**Theorem 9.55.** The wavelet functions  $\varphi(x), w_{j,k}(x)$  form an orthogonal system with respect to the inner product (9.124).

*Proof:* First, note that each wavelet  $w_{j,k}(x)$  is equal to  $+1$  on an interval of length  $2^{-j-1}$  and to  $-1$  on an adjacent interval of the same length. Therefore,

$$\langle w_{j,k}, \varphi \rangle = \int_0^1 w_{j,k}(x) dx = 0, \quad (9.134)$$

since the  $+1$  and  $-1$  contributions cancel each other. If two different wavelets  $w_{j,k}$  and  $w_{l,m}$  with, say  $j \leq l$ , have supports that are either disjoint, or just overlap at a single point, then their product  $w_{j,k}(x) w_{l,m}(x) \equiv 0$ , and so their inner product is clearly zero:

$$\langle w_{j,k}, w_{l,m} \rangle = \int_0^1 w_{j,k}(x) w_{l,m}(x) dx = 0.$$

Otherwise, except in the case when the two wavelets are identical, the support of  $w_{l,m}$  is entirely contained in an interval where  $w_{j,k}$  is constant, and so  $w_{j,k}(x) w_{l,m}(x) = \pm w_{l,m}(x)$ . Therefore, by (9.134),

$$\langle w_{j,k}, w_{l,m} \rangle = \int_0^1 w_{j,k}(x) w_{l,m}(x) dx = \pm \int_0^1 w_{l,m}(x) dx = 0.$$

Finally, we compute

$$\|\varphi\|^2 = \int_0^1 dx = 1, \quad \|w_{j,k}\|^2 = \int_0^1 w_{j,k}(x)^2 dx = 2^{-j}. \quad (9.135)$$

The second formula follows from the fact that  $|w_{j,k}(x)| = 1$  on an interval of length  $2^{-j}$  and is 0 elsewhere. *Q.E.D.*

The *wavelet series* of a signal  $f(x)$  is given by

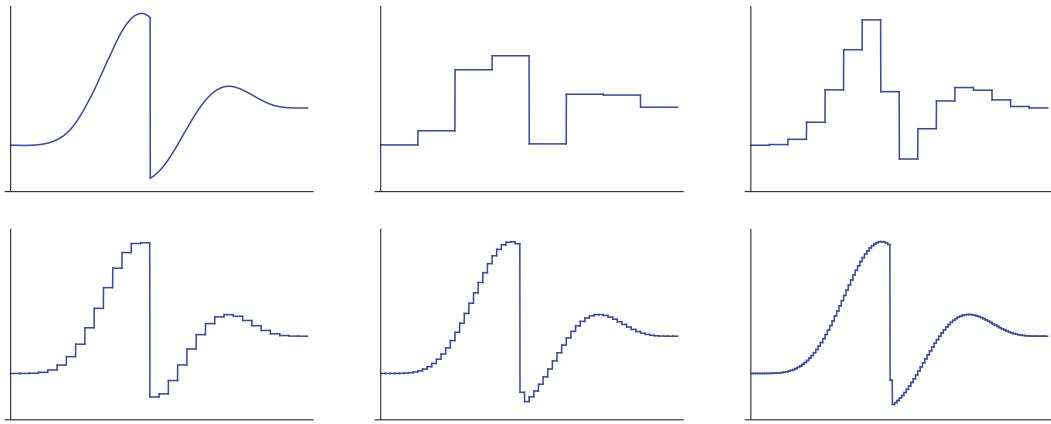
$$f(x) \sim c_0 \varphi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} c_{j,k} w_{j,k}(x). \quad (9.136)$$

Orthogonality implies that the wavelet coefficients  $c_0, c_{j,k}$  can be immediately computed using the standard inner product formula coupled with (9.135):

$$\begin{aligned} c_0 &= \frac{\langle f, \varphi \rangle}{\|\varphi\|^2} = \int_0^1 f(x) dx, \\ c_{j,k} &= \frac{\langle f, w_{j,k} \rangle}{\|w_{j,k}\|^2} = 2^j \int_{2^{-j}k}^{2^{-j}k+2^{-j-1}} f(x) dx - 2^j \int_{2^{-j}k+2^{-j-1}}^{2^{-j}(k+1)} f(x) dx. \end{aligned} \quad (9.137)$$

The convergence properties of the Haar wavelet series (9.136) are similar to those of Fourier series, [61, 77]; full details can be found [18, 88].

**Example 9.56.** In Figure 9.7, we plot the Haar expansions of the signal displayed in the first plot. The following plots show the partial sums for the Haar wavelet series (9.136) over  $j = 0, \dots, r$  with  $r = 2, 3, 4, 5, 6$ . Since the wavelets are themselves discontinuous, they do not have any difficulty converging to a discontinuous function. On the other hand, it takes quite a few wavelets to begin to accurately reproduce the signal — in the last plot, we are combining a total of  $2^6 = 64$  Haar wavelets.



**Figure 9.7.** Haar Wavelet Expansion.

## Exercises

♣ 9.7.1. Let  $f(x) = x$ . (a) Determine its Haar wavelet coefficients  $c_{j,k}$ . (b) Graph the partial sums  $s_r(x)$  of the Haar wavelet series (9.136) where  $j$  goes from 0 to  $r = 2, 5$ , and  $10$ . Compare your graphs with that of  $f$  and discuss what you observe. Is the series converging to the function? Can you prove this? (c) What is the maximal deviation  $\|f - s_r\|_\infty = \max\{|f(x) - s_r(x)| \mid 0 \leq x \leq 1\}$  for each of your partial sums?

♣ 9.7.2. Answer Exercise 9.7.1 for the functions

$$(a) x^2 - x, \quad (b) \cos \pi x, \quad (c) \begin{cases} e^{-x}, & 0 < x < \frac{1}{2}, \\ -e^{-x}, & \frac{1}{2} < x < 1. \end{cases}$$

♣ 9.7.3. In this exercise, we investigate the compression capabilities of the Haar wavelets. Let

$$f(x) = \begin{cases} -x, & 0 \leq x \leq \frac{1}{3}\pi, \\ x - \frac{2}{3}\pi, & \frac{1}{3}\pi \leq x \leq \frac{4}{3}\pi, \\ -x + 2\pi, & \frac{4}{3}\pi \leq x \leq 2\pi, \end{cases}$$

represent a signal defined on  $0 \leq x \leq 1$ . Let

$s_r(x)$  denote the  $n^{\text{th}}$  partial sum, from  $j = 0$  to  $r$ , of the Haar wavelet series (9.136).

(a) How many different Haar wavelet coefficients  $c_{j,k}$  appear in  $s_r(x)$ ? If our criterion for compression is that  $\|f - s_r\|_\infty < \varepsilon$ , how large do you need to choose  $r$  when  $\varepsilon = .1$ ?  $\varepsilon = .01$ ?  $\varepsilon = .001$ ? (b) Compare the Haar wavelet compression with the discrete Fourier method of Exercise 5.6.10.

♡ 9.7.4. (a) Explain why the wavelet expansion (9.136) defines a linear transformation on  $\mathbb{R}^n$  that takes a wavelet coefficient vector  $\mathbf{c} = (c_0, c_1, \dots, c_{n-1})^T$  to the corresponding sample vector  $\mathbf{f} = (f_0, f_1, \dots, f_{n-1})^T$ . (b) According to Theorem 7.5, the wavelet map must be given by matrix multiplication  $\mathbf{f} = W_n \mathbf{c}$  by a  $2 \times 2^n$  matrix  $W = W_n$ . Construct  $W_2$ ,  $W_3$  and  $W_4$ . (c) Prove that the columns of  $W_n$  are obtained as the values of the wavelet basis functions on the  $2^n$  sample intervals. (d) Prove that the columns of  $W_n$  are orthogonal. (e) Is  $W_n$  an orthogonal matrix? Find a formula for  $W_n^{-1}$ . (f) Explain why the wavelet transform is given by the linear map,  $\mathbf{c} = W_n^{-1} \mathbf{f}$ .

- ♠ 9.7.5. Test the noise removal features of the Haar wavelets by adding random noise to one of the functions in Exercises 9.7.1 and 9.7.2, computing the wavelet series, and then setting the high “frequency” modes to zero. What do you observe? Is this a reasonable denoising algorithm when compared with a Fourier method?
- 9.7.6. Write the Haar scaling function and mother wavelet as linear combinations of step functions.
- ◇ 9.7.7. Prove Lemma 9.54.

## Modern Wavelets

The main defect of the Haar wavelets is that they do not provide a very efficient means of representing even very simple functions — it takes quite a large number of wavelets to reproduce signals with any degree of precision. The reason for this is that the Haar wavelets are piecewise constant, and so even an affine function  $y = \alpha x + \beta$  requires many sample values, and hence a relatively extensive collection of Haar wavelets, to be accurately reproduced. In particular, compression and denoising algorithms based on Haar wavelets are either insufficiently precise or hopelessly inefficient, and hence of minor practical value.

For a long time it was thought that it was impossible to simultaneously achieve the requirements of localization, orthogonality and accurate reproduction of simple functions. The breakthrough came in 1988, when the Dutch mathematician Ingrid Daubechies produced the first examples of wavelet bases that realized all three basic criteria. Since then, wavelets have developed into a sophisticated and burgeoning industry with major impact on modern technology. Significant applications include compression, storage and recognition of fingerprints in the FBI’s data base, and the JPEG2000 image format, which, unlike earlier Fourier-based JPEG standards, incorporates wavelet technology in its image compression and reconstruction algorithms. In this section, we will present a brief outline of the basic ideas underlying Daubechies’ remarkable construction.

The recipe for any wavelet system involves two basic ingredients — a scaling function and a mother wavelet. The latter can be constructed from the scaling function by a prescription similar to that in (9.127), and therefore we first concentrate on the properties of the scaling function. The key requirement is that the scaling function must solve a *dilation equation* of the form

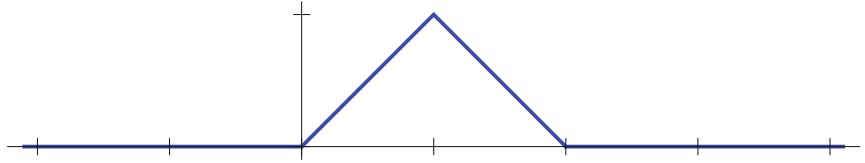
$$\varphi(x) = \sum_{k=0}^p c_k \varphi(2x - k) = c_0 \varphi(2x) + c_1 \varphi(2x - 1) + \cdots + c_p \varphi(2x - p) \quad (9.138)$$

for some collection of constants  $c_0, \dots, c_p$ . The dilation equation relates the function  $\varphi(x)$  to a finite linear combination of its compressed translates. The coefficients  $c_0, \dots, c_p$  are not arbitrary, since the properties of orthogonality and localization will impose certain rather stringent requirements.

**Example 9.57.** The Haar or box scaling function (9.125) satisfies the dilation equation (9.138) with  $c_0 = c_1 = 1$ , namely

$$\varphi(x) = \varphi(2x) + \varphi(2x - 1). \quad (9.139)$$

We recommend that you convince yourself of the validity of this identity before continuing.



**Figure 9.8.** The Hat Function.

**Example 9.58.** Another example of a scaling function is the *hat function*

$$\varphi(x) = \begin{cases} x, & 0 \leq x \leq 1, \\ 2-x, & 1 \leq x \leq 2, \\ 0, & \text{otherwise,} \end{cases} \quad (9.140)$$

graphed in Figure 9.8. The hat function satisfies the dilation equation

$$\varphi(x) = \frac{1}{2} \varphi(2x) + \varphi(2x-1) + \frac{1}{2} \varphi(2x-2), \quad (9.141)$$

which is (9.138) with  $c_0 = \frac{1}{2}$ ,  $c_1 = 1$ ,  $c_2 = \frac{1}{2}$ . Again, the reader should be able to check this identity by hand.

The dilation equation (9.138) is a kind of *functional equation*, and, as such, is not so easy to solve. Indeed, the mathematics of functional equations remains much less well developed than that of differential equations or integral equations. Even to prove that (nonzero) solutions exist is a nontrivial analytical problem. Since we already know two explicit examples, let us defer the discussion of solution techniques until we understand how the dilation equation can be used to construct a wavelet basis.

Given a solution to the dilation equation, we define the *mother wavelet* to be

$$\begin{aligned} w(x) &= \sum_{k=0}^p (-1)^k c_{p-k} \varphi(2x-k) \\ &= c_p \varphi(2x) - c_{p-1} \varphi(2x-1) + c_{p-2} \varphi(2x-2) + \dots \pm c_0 \varphi(2x-p). \end{aligned} \quad (9.142)$$

This formula directly generalizes the Haar wavelet relation (9.127), in light of its dilation equation (9.139). The daughter wavelets are then all found, as in the Haar basis, by iteratively compressing and translating the mother wavelet:

$$w_{j,k}(x) = w(2^j x - k). \quad (9.143)$$

In the general framework, we do not necessarily restrict our attention to the interval  $[0, 1]$ , and so  $j$  and  $k$  can, in principle, be arbitrary integers.

Let us investigate what sort of conditions should be imposed on the dilation coefficients  $c_0, \dots, c_p$  in order that we obtain a viable wavelet basis by this construction. First, localization of the wavelets requires that the scaling function have bounded support, and so  $\varphi(x) \equiv 0$  when  $x$  lies outside some bounded interval  $[a, b]$ . Integrating both sides of (9.138) produces

$$\int_a^b \varphi(x) dx = \int_{-\infty}^{\infty} \varphi(x) dx = \sum_{k=0}^p c_k \int_{-\infty}^{\infty} \varphi(2x-k) dx. \quad (9.144)$$

Performing the change of variables  $y = 2x - k$ , with  $dx = \frac{1}{2} dy$ , we obtain

$$\int_{-\infty}^{\infty} \varphi(2x-k) dx = \frac{1}{2} \int_{-\infty}^{\infty} \varphi(y) dy = \frac{1}{2} \int_a^b \varphi(x) dx, \quad (9.145)$$

where we revert to  $x$  as our (dummy) integration variable. We substitute this result back into (9.144). Assuming that  $\int_a^b \varphi(x) dx \neq 0$ , we discover that the dilation coefficients must satisfy

$$c_0 + \cdots + c_p = 2. \quad (9.146)$$

The second condition we require is orthogonality of the wavelets. For simplicity, we only consider the standard  $L^2$  inner product<sup>†</sup>

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) dx.$$

It turns out that the orthogonality of the complete wavelet system is guaranteed once we know that the scaling function  $\varphi(x)$  is orthogonal to all its integer translates:

$$\langle \varphi(x), \varphi(x - m) \rangle = \int_{-\infty}^{\infty} \varphi(x) \varphi(x - m) dx = 0 \quad \text{for all } m \neq 0. \quad (9.147)$$

We first note the formula

$$\begin{aligned} \langle \varphi(2x - k), \varphi(2x - l) \rangle &= \int_{-\infty}^{\infty} \varphi(2x - k) \varphi(2x - l) dx \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \varphi(x) \varphi(x + k - l) dx = \frac{1}{2} \langle \varphi(x), \varphi(x + k - l) \rangle \end{aligned} \quad (9.148)$$

follows from the same change of variables  $y = 2x - k$  used in (9.145). Therefore, since  $\varphi$  satisfies the dilation equation (9.138), we have

$$\begin{aligned} \langle \varphi(x), \varphi(x - m) \rangle &= \left\langle \sum_{j=0}^p c_j \varphi(2x - j), \sum_{k=0}^p c_k \varphi(2x - 2m - k) \right\rangle \\ &= \sum_{j,k=0}^p c_j c_k \langle \varphi(2x - j), \varphi(2x - 2m - k) \rangle = \frac{1}{2} \sum_{j,k=0}^p c_j c_k \langle \varphi(x), \varphi(x + j - 2m - k) \rangle. \end{aligned} \quad (9.149)$$

If we require orthogonality (9.147) of all the integer translates of  $\varphi$ , then the left-hand side of this identity will be 0 unless  $m = 0$ , while only the summands with  $j = 2m + k$  will be nonzero on the right. Therefore, orthogonality requires that

$$\sum_{0 \leq k \leq p-2m} c_{2m+k} c_k = \begin{cases} 2, & m = 0, \\ 0, & m \neq 0. \end{cases} \quad (9.150)$$

The algebraic equations (9.146, 150) for the dilation coefficients are the key requirements for the construction of an orthogonal wavelet basis.

For example, if we have just two nonzero coefficients  $c_0, c_1$ , then (9.146, 150) reduce to

$$c_0 + c_1 = 2, \quad c_0^2 + c_1^2 = 2,$$

and so  $c_0 = c_1 = 1$  is the only solution, resulting in the Haar dilation equation (9.139). If we have three coefficients  $c_0, c_1, c_2$ , then (9.146), (9.150) require

$$c_0 + c_1 + c_2 = 2, \quad c_0^2 + c_1^2 + c_2^2 = 2, \quad c_0 c_2 = 0.$$

---

<sup>†</sup> In all instances, the functions have bounded support, and so the inner product integral can be reduced to an integral over a finite interval where both  $f$  and  $g$  are nonzero.

Thus either  $c_2 = 0$ ,  $c_0 = c_1 = 1$ , and we are back to the Haar case, or  $c_0 = 0$ ,  $c_1 = c_2 = 1$ , and the resulting dilation equation is a simple reformulation of the Haar case. In particular, the hat function (9.140) does *not* give rise to orthogonal wavelets.

The remarkable fact, discovered by Daubechies, is that there *is* a nontrivial solution for four (and, indeed, any even number) of nonzero coefficients  $c_0, c_1, c_2, c_3$ . The basic equations (9.146), (9.150) require

$$c_0 + c_1 + c_2 + c_3 = 2, \quad c_0^2 + c_1^2 + c_2^2 + c_3^2 = 2, \quad c_0 c_2 + c_1 c_3 = 0. \quad (9.151)$$

The particular values

$$c_0 = \frac{1 + \sqrt{3}}{4}, \quad c_1 = \frac{3 + \sqrt{3}}{4}, \quad c_2 = \frac{3 - \sqrt{3}}{4}, \quad c_3 = \frac{1 - \sqrt{3}}{4}, \quad (9.152)$$

solve (9.151). These coefficients correspond to the *Daubechies dilation equation*

$$\varphi(x) = \frac{1 + \sqrt{3}}{4} \varphi(2x) + \frac{3 + \sqrt{3}}{4} \varphi(2x - 1) + \frac{3 - \sqrt{3}}{4} \varphi(2x - 2) + \frac{1 - \sqrt{3}}{4} \varphi(2x - 3). \quad (9.153)$$

A nonzero solution of bounded support to this remarkable functional equation will give rise to a scaling function  $\varphi(x)$ , a mother wavelet

$$w(x) = \frac{1 - \sqrt{3}}{4} \varphi(2x) - \frac{3 - \sqrt{3}}{4} \varphi(2x - 1) + \frac{3 + \sqrt{3}}{4} \varphi(2x - 2) - \frac{1 + \sqrt{3}}{4} \varphi(2x - 3), \quad (9.154)$$

and then, by compression and translation (9.143), the complete system of orthogonal wavelets  $w_{j,k}(x)$ .

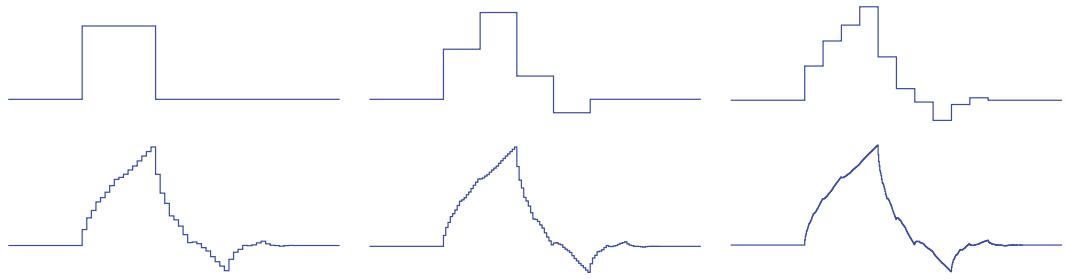
Before explaining how to solve the Daubechies dilation equation, let us complete the proof of orthogonality. It is easy to see that, by translation invariance, since  $\varphi(x)$  and  $\varphi(x - m)$  are orthogonal whenever  $m \neq 0$ , so are  $\varphi(x - k)$  and  $\varphi(x - l)$  for all  $k \neq l$ . Next we prove orthogonality of  $\varphi(x - m)$  and  $w(x)$ :

$$\begin{aligned} \langle w(x), \varphi(x - m) \rangle &= \left\langle \sum_{j=0}^p (-1)^{j+1} c_j \varphi(2x - 1 + j), \sum_{k=0}^p c_k \varphi(2x - 2m - k) \right\rangle \\ &= \sum_{j,k=0}^p (-1)^{j+1} c_j c_k \langle \varphi(2x - 1 + j), \varphi(2x - 2m - k) \rangle \\ &= \frac{1}{2} \sum_{j,k=0}^p (-1)^{j+1} c_j c_k \langle \varphi(x), \varphi(x - 1 + j - 2m - k) \rangle, \end{aligned}$$

using (9.148). By orthogonality (9.147) of the translates of  $\varphi$ , the only summands that are nonzero are those for which  $j = 2m + k + 1$ ; the resulting coefficient of  $\|\varphi(x)\|^2$  is

$$\sum_k (-1)^k c_{1-2m-k} c_k = 0,$$

where the sum is over all  $0 \leq k \leq p$  such that  $0 \leq 1 - 2m - k \leq p$ . Each term in the sum appears twice, with opposite signs, and hence the result is always zero — no matter what the coefficients  $c_0, \dots, c_p$  are! The proof of orthogonality of the translates  $w(x - m)$  of the mother wavelet, along with all her wavelet descendants  $w(2^j x - k)$ , relies on a similar argument, and the details are left as an exercise for the reader.



**Figure 9.9.** Approximating the Daubechies Wavelet.

### Solving the Dilation Equation

Let us next discuss how to solve the dilation equation (9.138). The solution we are after does not have an elementary formula, and we require a slightly sophisticated approach to recover it. The key observation is that (9.138) has the form of a fixed point equation

$$\varphi = F[\varphi],$$

not in ordinary Euclidean space, but in an infinite-dimensional function space. With luck, the fixed point (or, more correctly, fixed function) will be stable, and so starting with a suitable initial guess  $\varphi_0(x)$ , the successive iterates

$$\varphi_{n+1} = F[\varphi_n]$$

will converge to the desired solution:  $\varphi_n(x) \rightarrow \varphi(x)$ . In detail, the iterative version of the dilation equation (9.138) reads

$$\varphi_{n+1}(x) = \sum_{k=0}^p c_k \varphi_n(2x - k), \quad n = 0, 1, 2, \dots. \quad (9.155)$$

Before attempting to prove convergence of this iterative procedure to the Daubechies scaling function, let us experimentally investigate what happens.

A reasonable choice for the initial guess might be the Haar scaling or box function

$$\varphi_0(x) = \begin{cases} 1, & 0 < t \leq 1. \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 9.9 we graph the subsequent iterates  $\varphi_1(x), \varphi_2(x), \varphi_4(x), \varphi_5(x), \varphi_7(x)$ . There clearly appears to be convergence to some function  $\varphi(x)$ , although the final result looks a little bizarre. Bolstered by this preliminary experimental evidence, we can now try to prove convergence of the iterative scheme. This turns out to be true; a fully rigorous proof relies on the Fourier transform, and can be found in [18].

**Theorem 9.59.** The functions  $\varphi_n(x)$  defined by the iterative functional equation (9.155) converge uniformly to a continuous function  $\varphi(x)$ , called the *Daubechies scaling function*.

Once we have established convergence, we are now able to verify that the scaling function and consequential system of wavelets form an orthogonal system of functions.

**Proposition 9.60.** All integer translates  $\varphi(x - k)$ , for  $k \in \mathbb{Z}$  of the Daubechies scaling function, and all wavelets  $w_{j,k}(x) = w(2^j x - k)$ ,  $j \geq 0$ , are mutually orthogonal functions with respect to the  $L^2$  inner product. Moreover,  $\|\varphi\|^2 = 1$ , while  $\|w_{j,k}\|^2 = 2^{-j}$ .

*Proof:* As noted earlier, the orthogonality of the entire wavelet system will follow once we know the orthogonality (9.147) of the scaling function and its integer translates. We use induction to prove that this holds for all the iterates  $\varphi_n(x)$ , and so, in view of uniform convergence, the limiting scaling function also satisfies this property. We already know that the orthogonality property holds for the Haar scaling function  $\varphi_0(x)$ . To demonstrate the induction step, we repeat the computation in (9.149), but now the left-hand side is  $\langle \varphi_{n+1}(x), \varphi_{n+1}(x - m) \rangle$ , while all other terms involve the previous iterate  $\varphi_n$ . In view of the algebraic constraints (9.150) on the wavelet coefficients and the induction hypothesis, we deduce that  $\langle \varphi_{n+1}(x), \varphi_{n+1}(x - m) \rangle = 0$  whenever  $m \neq 0$ , while when  $m = 0$ ,  $\|\varphi_{n+1}\|^2 = \|\varphi_n\|^2$ . Since  $\|\varphi_0\| = 1$ , we further conclude that all the iterates, and hence the limiting scaling function, all have unit  $L^2$  norm. The proof of the formula for the norms of the mother and daughter wavelets is left for Exercise 9.7.19. *Q.E.D.*

In practical computations, the limiting procedure for constructing the scaling function is not so convenient, and an alternative means of computing its values is employed. The starting point is to determine its values at integer points. First, the initial box function has values  $\varphi_0(m) = 0$  for all integers  $m \in \mathbb{Z}$  except  $\varphi_0(1) = 1$ . The iterative functional equation (9.155) will then produce the values of the iterates  $\varphi_n(m)$  at integer points  $m \in \mathbb{Z}$ . A simple induction will convince you that  $\varphi_n(m) = 0$  except for  $m = 1$  and  $m = 2$ , and, therefore, by (9.155),

$$\varphi_{n+1}(1) = \frac{3 + \sqrt{3}}{4} \varphi_n(1) + \frac{1 + \sqrt{3}}{4} \varphi_n(2), \quad \varphi_{n+1}(2) = \frac{1 - \sqrt{3}}{4} \varphi_n(1) + \frac{3 - \sqrt{3}}{4} \varphi_n(2),$$

since all other terms are 0. This has the form of a linear iterative system

$$\mathbf{v}^{(n+1)} = A \mathbf{v}^{(n)} \quad (9.156)$$

with coefficient matrix

$$A = \begin{pmatrix} \frac{3 + \sqrt{3}}{4} & \frac{1 + \sqrt{3}}{4} \\ \frac{1 - \sqrt{3}}{4} & \frac{3 - \sqrt{3}}{4} \end{pmatrix} \quad \text{and where} \quad \mathbf{v}^{(n)} = \begin{pmatrix} \varphi_n(1) \\ \varphi_n(2) \end{pmatrix}.$$

As we know, the solution to such an iterative system is specified by the eigenvalues and eigenvectors of the coefficient matrix, which are

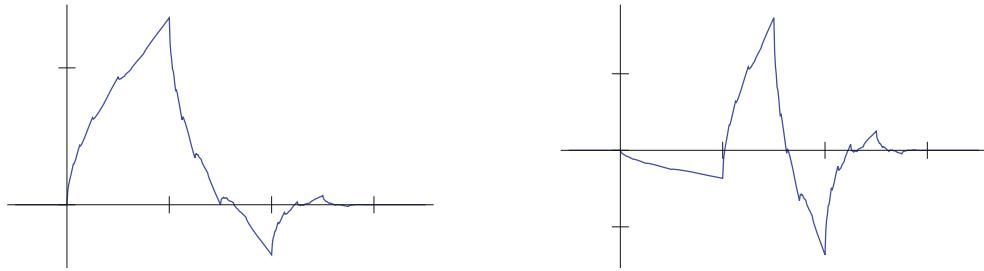
$$\lambda_1 = 1, \quad \mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{4} \\ \frac{1-\sqrt{3}}{4} \end{pmatrix}, \quad \lambda_2 = \frac{1}{2}, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We write the initial condition as a linear combination of the eigenvectors

$$\mathbf{v}^{(0)} = \begin{pmatrix} \varphi_0(1) \\ \varphi_0(2) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \mathbf{v}_1 - \frac{1 - \sqrt{3}}{2} \mathbf{v}_2.$$

The solution is

$$\mathbf{v}^{(n)} = A^n \mathbf{v}^{(0)} = 2 A^n \mathbf{v}_1 - \frac{1 - \sqrt{3}}{2} A^n \mathbf{v}_2 = 2 \mathbf{v}_1 - \frac{1}{2^n} \frac{1 - \sqrt{3}}{2} \mathbf{v}_2.$$



**Figure 9.10.** The Daubechies Scaling Function and Mother Wavelet.

The limiting vector

$$\begin{pmatrix} \varphi(1) \\ \varphi(2) \end{pmatrix} = \lim_{n \rightarrow \infty} \mathbf{v}^{(n)} = 2\mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{2} \\ \frac{2}{1-\sqrt{3}} \end{pmatrix}$$

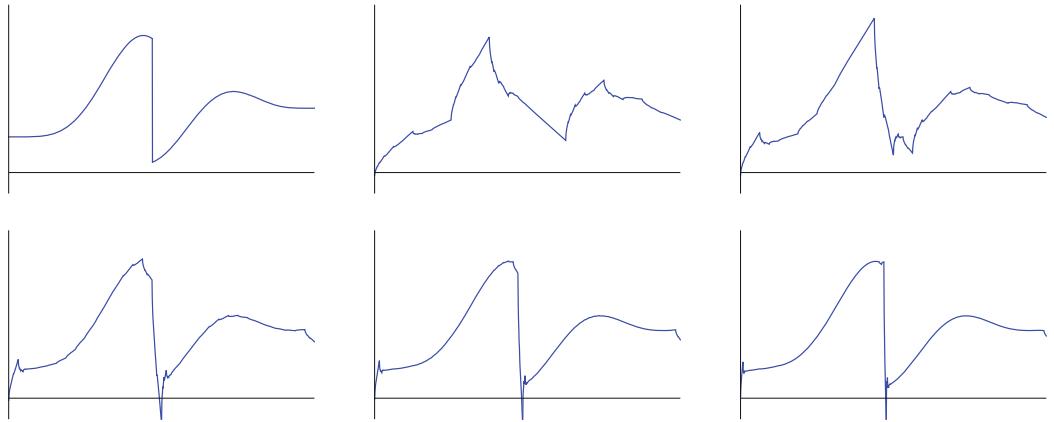
gives the desired values of the scaling function:

$$\begin{aligned} \varphi(1) &= \frac{1+\sqrt{3}}{2} = 1.366025\dots, & \varphi(2) &= \frac{1-\sqrt{3}}{2} = -.366025\dots, \\ \varphi(m) &= 0, & \text{for all } m &\neq 1, 2. \end{aligned} \quad (9.157)$$

With this in hand, the Daubechies dilation equation (9.153) then prescribes the function values  $\varphi(\frac{1}{2}m)$  at all half-integers, because if  $x = \frac{1}{2}m$ , then  $2x - k = m - k$  is an integer. Once we know its values at the half-integers, we can reuse equation (9.153) to give its values at quarter-integers  $\frac{1}{4}m$ . Continuing onward, we determine the values of  $\varphi(x)$  at all *dyadic points*, meaning rational numbers of the form  $x = m/2^j$  for  $m, j \in \mathbb{Z}$ . Continuity will then prescribe its value at all other  $x \in \mathbb{R}$  since  $x$  can be written as the limit of dyadic numbers  $x_n$  — namely those obtained by truncating its binary (base 2) expansion at the  $n^{\text{th}}$  digit beyond the decimal (or, rather “binary”) point. But, in practice, this latter step is unnecessary, since all computers are ultimately based on the binary number system, and so only dyadic numbers actually reside in a computer’s memory. Thus, there is no real need to determine the value of  $\varphi$  at non-dyadic points.

The preceding scheme was used to produce the graphs of the Daubechies scaling function in [Figure 9.10](#). It is a continuous, but non-differentiable, function — and its graph has a very jagged, fractal-like appearance when viewed at close range. The Daubechies scaling function is, in fact, a close relative of the famous example of a continuous, nowhere differentiable function originally due to Weierstrass, [\[42, 53\]](#), whose construction also relies on a similar scaling argument.

Given the values of the Daubechies scaling function on a sufficiently dense set of dyadic points, the consequential values of the mother wavelet are given by formula (9.154). Note that  $\text{supp } \varphi = \text{supp } w = [0, 3]$ . The daughter wavelets are then found by the usual compression and translation procedure (9.143).



**Figure 9.11.** Daubechies Wavelet Expansion.

The Daubechies wavelet expansion of a function whose support is contained in<sup>†</sup>  $[0, 1]$  is then given by

$$f(x) \sim c_0 \varphi(x) + \sum_{j=0}^{\infty} \sum_{k=-2}^{2^j-1} c_{j,k} w_{j,k}(x). \quad (9.158)$$

The inner summation begins at  $k = -2$  so as to include *all* the wavelet offspring  $w_{j,k}$  whose supports have a nontrivial intersection with the interval  $[0, 1]$ . The wavelet coefficients  $c_0, c_{j,k}$  are computed by the usual orthogonality formula

$$\begin{aligned} c_0 &= \langle f, \varphi \rangle = \int_0^3 f(x) \varphi(x) dx, \\ c_{j,k} &= \langle f, w_{j,k} \rangle = 2^j \int_{2^{-j}k}^{2^{-j}(k+3)} f(x) w_{j,k}(x) dx = \int_0^3 f(2^{-j}(x+k)) w(x) dx, \end{aligned} \quad (9.159)$$

where we agree that  $f(x) = 0$  whenever  $x < 0$  or  $x > 1$ . In practice, one employs a numerical integration procedure, e.g., the trapezoid rule, based on dyadic points to speedily evaluate the integrals (9.159). A proof of completeness of the resulting wavelet basis functions can be found in [18]. Compression and denoising algorithms based on retaining only low-frequency modes proceed as in Section 5.6, and are left as projects for the motivated reader to implement.

**Example 9.61.** In Figure 9.11, we plot the Daubechies wavelet expansions of the same signal for Example 9.56. The first plot is the original signal, and the following show the partial sums of (9.158) over  $j = 0, \dots, r$  with  $r = 2, 3, 4, 5, 6$ . Unlike the Haar expansion, the Daubechies wavelets do exhibit a nonuniform Gibbs phenomenon, where the expansion noticeably overshoots near the discontinuity, [61], which can be observed at the interior discontinuity as well as the endpoints, since the function is set to 0 outside the interval

<sup>†</sup> For functions with larger support, one should include additional terms in the expansion corresponding to further translates of the wavelets so as to cover the entire support of the function. Alternatively, one can translate and rescale  $x$  to fit the function's support inside  $[0, 1]$ .

$[0, 1]$ . Indeed, the Daubechies wavelets are continuous, and so cannot converge uniformly to a discontinuous function.

## Exercises

- ♠ 9.7.8. Answer Exercises 9.7.1 and 9.7.2 using the Daubechies wavelets instead of the Haar wavelets. Do you see any improvement in your approximations? Discuss the advantages and disadvantages of both in light of these examples.
- ♠ 9.7.9. Answer Exercise 9.7.3 using the Daubechies wavelets to compress the data. Compare your results.
- ◇ 9.7.10. Verify formulas (9.139) and (9.141).
- 9.7.11. Prove that the most general solution to the functional equation  $\varphi(x) = 2\varphi(2x)$  is  $\varphi(x) = f(\log_2 x)/x$  where  $f(z+1) = f(z)$  is any 1 periodic function.
- ◇ 9.7.12. Consider the dilation equation (9.138) with  $c_0 = 0$ ,  $c_1 = c_2 = 1$ , so  $\varphi(x) = \varphi(2x-1) + \varphi(2x-2)$ . Prove that  $\psi(x) = \varphi(x+1)$  satisfies the Haar dilation equation (9.139). Generalize this result to prove that we can always, without loss of generality, assume that  $c_0 \neq 0$  in the general dilation equation (9.138).
- 9.7.13. Prove that a cubic  $B$  spline, as defined in Exercise 5.5.76, solves the dilation equation (9.138) for  $c_0 = c_4 = \frac{1}{8}$ ,  $c_1 = c_3 = \frac{1}{2}$ ,  $c_2 = \frac{3}{4}$ .
- 9.7.14. Explain why the scaling function  $\varphi(x)$  and the mother wavelet  $w(x)$  have the same support:  $\text{supp } \varphi = \text{supp } w$ .
- 9.7.15. Prove that (9.147) implies  $\langle \varphi(x-l), \varphi(x-m) \rangle = 0$  for all  $l \neq m$ .
- ◇ 9.7.16. Let  $\varphi(x)$  be any scaling function,  $w(x)$  the corresponding mother wavelet and  $w_{j,k}(x)$  the wavelet descendants. Prove that (a)  $\|\varphi\| = \|w\|$ . (b)  $\|w_{j,k}\| = 2^{-j} \|\varphi\|$ .
- ◇ 9.7.17. (a) Prove that the scaling function  $\varphi(x)$  and the mother wavelet  $w(x)$  are orthogonal. (b) Prove that the integer translates  $w(x-m)$  of the mother wavelet are mutually orthogonal. (c) Prove orthogonality of all the wavelet offspring  $w_{j,k}(x)$ .
- 9.7.18. Find the values of the Daubechies scaling function  $\varphi(x)$  and mother wavelet  $w(x)$  at  $x =$  (a)  $\frac{1}{2}$ , (b)  $\frac{1}{4}$ , (c)  $\frac{5}{16}$ .
- ◇ 9.7.19. Prove the formulas in Proposition 9.60 for the norms of the mother and daughter wavelets.
- ♠ 9.7.20. Write a computer program to zoom in on the Daubechies scaling function and discuss what you see.
- 9.7.21. *True or false:* The iterative system (9.156) is a Markov process.
- ◇ 9.7.22. Let  $\varphi(x)$  satisfy the Daubechies scaling equation (9.153). Prove that if  $\varphi(i) \neq 0$  for any  $i \leq 0$  or  $i \geq p$ , then  $\text{supp } \varphi$  is unbounded.
- 9.7.23. (a) Use (9.142) to construct the “mother wavelet” corresponding to the hat function (9.140). (b) Is the hat function orthogonal to the mother wavelet? (c) Is the hat function orthogonal to its integer translates?
- 9.7.24. Prove that a real number  $x$  is dyadic if and only if its binary (base 2) expansion terminates, i.e., is eventually all zeros.
- 9.7.25. Find dyadic approximations, with error at most  $2^{-8}$ , to  
 (a)  $\frac{3}{4}$ , (b)  $\frac{1}{3}$ , (c)  $\sqrt{2}$ , (d)  $e$ , (e)  $\pi$ .



# Chapter 10

# Dynamics

In this chapter, we turn our attention to continuous dynamical systems, which are governed by first and second order linear systems of ordinary differential equations. Such systems, whose unvarying equilibria were the subject of Chapter 6, include the dynamical motions of mass–spring chains and structures, and the time-varying voltages and currents in simple electrical circuits. Dynamics of continuous media, including fluids, solids, and gases, are modeled by infinite-dimensional dynamical systems described by partial differential equations, [61, 79], and will not be treated in this text, nor will we venture into the vastly more complicated realm of nonlinear dynamics, [34, 41].

Chapter 8 developed the basic mathematical tools — eigenvalues and eigenvectors — used in the analysis of linear systems of ordinary differential equations. For a first order system, the resulting *eigensolutions* describe the basic modes of exponential growth, decay, or periodic behavior. In particular, the stability properties of an equilibrium solution are (mostly) determined by the eigenvalues. Most of the phenomenology inherent in linear dynamics can already be observed in the two-dimensional situation, and we devote Section 10.3 to a complete description of first order planar linear systems. In Section 10.4, we re-interpret the solution to a first order system in terms of the matrix exponential, which is defined by analogy with the usual scalar exponential function. Matrix exponentials are particularly effective for solving inhomogeneous or forced linear systems, and also appear in applications to geometry, computer graphics and animation, theoretical physics, and mechanics.

As a consequence of Newton’s laws of motion, mechanical vibrations are modeled by second order dynamical systems. For stable configurations with no frictional damping, the eigensolutions constitute the system’s normal modes, each periodically vibrating with its associated natural frequency. The full dynamics is obtained by linear superposition of the periodic normal modes, but the resulting solution is, typically, no longer periodic. Such quasi-periodic motion may seem quite chaotic — even though mathematically it is merely a combination of finitely many simple periodic solutions. When subjected to an external periodic forcing, the system usually remains in a quasi-periodic motion that superimposes a periodic response onto its own internal vibrations. However, attempting to force the system at one of its natural frequencies, as prescribed by its eigenvalues, may induce a resonant vibration, of progressively unbounded amplitude, resulting in a catastrophic breakdown of the physical apparatus. In contrast, frictional effects, depending on first order derivatives/velocities, serve to damp out the quasi-periodic vibrations and similarly help mitigate the dangers of resonance.

## 10.1 Basic Solution Techniques

Our initial focus will be on systems

$$\frac{d\mathbf{u}}{dt} = A \mathbf{u} \quad (10.1)$$

consisting of  $n$  first order linear ordinary differential equations in the  $n$  unknowns  $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))^T \in \mathbb{R}^n$ . In an *autonomous* system, the time variable does not appear explicitly, and so the coefficient matrix  $A$ , of size  $n \times n$ , is a constant real<sup>†</sup> matrix. Non-autonomous systems, in which  $A(t)$  is time-dependent, are considerably more difficult to analyze, and we refer the reader to a more advanced text such as [36].

As we saw in Section 8.1, a vector-valued exponential function

$$\mathbf{u}(t) = e^{\lambda t} \mathbf{v},$$

in which  $\lambda$  is a constant scalar and  $\mathbf{v}$  a constant vector, describes a solution to (10.1) if and only if

$$A\mathbf{v} = \lambda\mathbf{v}.$$

Hence, assuming  $\mathbf{v} \neq \mathbf{0}$ , the scalar  $\lambda$  must be an eigenvalue of  $A$ , and  $\mathbf{v}$  the corresponding eigenvector. The resulting exponential function will be called an *eigensolution* of the linear system. Since the system is linear and homogeneous, linear superposition allows us to combine the basic eigensolutions to form more general solutions.

If the coefficient matrix  $A$  is complete (diagonalizable), then, by definition, its eigenvectors  $\mathbf{v}_1, \dots, \mathbf{v}_n$  form a basis. The corresponding eigensolutions

$$\mathbf{u}_1(t) = e^{\lambda_1 t} \mathbf{v}_1, \quad \dots \quad \mathbf{u}_n(t) = e^{\lambda_n t} \mathbf{v}_n,$$

will form a basis for the solution space to the system. Hence, the general solution to a first order linear system with complete coefficient matrix has the form

$$\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \dots + c_n \mathbf{u}_n(t) = c_1 e^{\lambda_1 t} \mathbf{v}_1 + \dots + c_n e^{\lambda_n t} \mathbf{v}_n, \quad (10.2)$$

where  $c_1, \dots, c_n$  are constants, which are uniquely prescribed by the initial conditions

$$\mathbf{u}(t_0) = \mathbf{u}_0. \quad (10.3)$$

This all follows from the basic existence and uniqueness theorem for ordinary differential equations, which will be discussed shortly.

**Example 10.1.** Let us solve the coupled pair of ordinary differential equations

$$\frac{du}{dt} = 3u + v, \quad \frac{dv}{dt} = u + 3v.$$

We first write the system in matrix form (10.1) with unknown  $\mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}$  and coefficient matrix  $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ . According to Example 8.5, the eigenvalues and eigenvectors of  $A$  are

$$\lambda_1 = 4, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = 2, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Both eigenvalues are simple, and so  $A$  is a complete matrix. The resulting eigensolutions

$$\mathbf{u}_1(t) = e^{4t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} e^{4t} \\ e^{4t} \end{pmatrix}, \quad \mathbf{u}_2(t) = e^{2t} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -e^{2t} \\ e^{2t} \end{pmatrix},$$

---

<sup>†</sup> Extending the solution techniques to complex systems with complex coefficient matrices is straightforward, but will not be treated here.

form a basis of the solution space, and so the general solution is a linear combination

$$\mathbf{u}(t) = c_1 e^{4t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2 e^{2t} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 e^{4t} - c_2 e^{2t} \\ c_1 e^{4t} + c_2 e^{2t} \end{pmatrix}, \quad \text{hence} \quad \begin{aligned} u(t) &= c_1 e^{4t} - c_2 e^{2t}, \\ v(t) &= c_1 e^{4t} + c_2 e^{2t}, \end{aligned}$$

in which  $c_1, c_2$  are arbitrary constants.

## The Phase Plane

As noted above, a wide variety of physical systems are modeled by second order ordinary differential equations. Your first course on ordinary differential equations, e.g., [7, 22], covered the basic solution technique for constant coefficient scalar equations, which we quickly review in the context of an example.

**Example 10.2.** To solve the homogeneous ordinary differential equation

$$\frac{d^2u}{dt^2} + \frac{du}{dt} - 6u = 0, \quad (10.4)$$

we begin with the exponential ansatz<sup>†</sup>

$$u(t) = e^{\lambda t},$$

where the constant factor  $\lambda$  is to be determined. Substituting into the differential equation leads immediately to the *characteristic equation*

$$\lambda^2 + \lambda - 6 = 0, \quad \text{with roots} \quad \lambda_1 = 2, \quad \lambda_2 = -3.$$

Therefore,  $e^{2t}$  and  $e^{-3t}$  are individual solutions. Since the equation is of second order, Theorem 7.34 implies that they form a basis for the two-dimensional solution space, and hence the general solution can be written as a linear combination

$$u(t) = c_1 e^{2t} + c_2 e^{-3t}, \quad (10.5)$$

where  $c_1, c_2$  are arbitrary constants.

There is a standard trick to convert a second order equation

$$\frac{d^2u}{dt^2} + \alpha \frac{du}{dt} + \beta u = 0 \quad (10.6)$$

into a first order system. One introduces the so-called *phase plane variables*<sup>‡</sup>

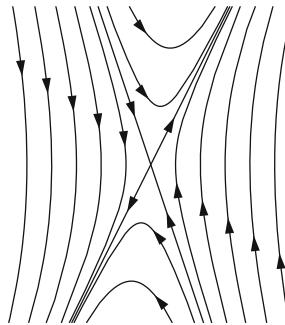
$$u_1 = u, \quad u_2 = \dot{u} = \frac{du}{dt}. \quad (10.7)$$

Assuming  $\alpha, \beta$  are constants, the phase plane variables satisfy

$$\frac{du_1}{dt} = \frac{du}{dt} = u_2, \quad \frac{du_2}{dt} = \frac{d^2u}{dt^2} = -\beta u - \alpha \frac{du}{dt} = -\beta u_1 - \alpha u_2.$$

<sup>†</sup> See the footnote on p. 379 for an explanation of the term “ansatz”, a.k.a. “inspired guess”.

<sup>‡</sup> We will often use dots as a shorthand notation for time derivatives.



**Figure 10.1.** Phase Portrait of  $\dot{u}_1 = u_2$ ,  $\dot{u}_2 = 6u_1 - u_2$ .

In this manner, the second order equation (10.6) is converted into a first order system

$$\dot{\mathbf{u}} = A\mathbf{u}, \quad \text{where} \quad \mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 \\ -\beta & -\alpha \end{pmatrix}. \quad (10.8)$$

Every solution  $u(t)$  to the second order equation yields a solution  $\mathbf{u}(t) = (u(t), \dot{u}(t))^T$  to the first order system (10.8), whose second component is merely its time derivative. Conversely, if  $\mathbf{u}(t) = (u_1(t), u_2(t))^T$  is any solution to (10.8), then its first component  $u(t) = u_1(t)$  defines a solution to the original scalar equation (10.6). We conclude that the two are completely equivalent, in the sense that solving one will immediately resolve the other.

The variables  $(u_1, u_2)^T = (u, \dot{u})^T$  serve as coordinates in the *phase plane*  $\mathbb{R}^2$ . The solutions  $\mathbf{u}(t)$  parameterize curves in the phase plane, known as the solution *trajectories* or *orbits*. In particular, the equilibrium solution  $\mathbf{u}(t) \equiv \mathbf{0}$  remains fixed at the origin, and so its trajectory is a single point. Assuming  $\beta \neq 0$ , every other solution describes a genuine curve, whose tangent direction  $\dot{\mathbf{u}} = d\mathbf{u}/dt$  at a point  $\mathbf{u}$  is prescribed by the right-hand side of the differential equation, namely  $\dot{\mathbf{u}} = A\mathbf{u}$ . The collection of all possible solution trajectories is called the *phase portrait* of the system. An important fact is that, in an autonomous first order system, the *phase plane trajectories never cross*. This striking property, which is also valid for nonlinear systems, is a consequence of the uniqueness properties of solutions, [7, 36]. Thus, the phase portrait consists of a family of non-intersecting curves that, when combined with the equilibrium points, fill out the entire phase plane. The direction of motion along a trajectory will be indicated graphically by a small arrow; nearby trajectories are all traversed in the same direction. The one feature that is not so easily pictured in the phase portrait is the continuously varying speed at which the solution moves along its trajectory. Plotting this requires a more complicated three-dimensional diagram using time as the third coordinate.

**Example 10.2 (continued).** For the second order equation (10.4), the equivalent phase plane system is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ 6 & -1 \end{pmatrix} \mathbf{u}, \quad \text{or, in full detail,} \quad \begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= 6u_1 - u_2. \end{aligned} \quad (10.9)$$

Our previous solution formula (10.5) implies that the solution to the phase plane system (10.9) is given by

$$u_1(t) = u(t) = c_1 e^{2t} + c_2 e^{-3t}, \quad u_2(t) = \frac{du}{dt} = 2c_1 e^{2t} - 3c_2 e^{-3t},$$

and hence

$$\mathbf{u}(t) = \begin{pmatrix} c_1 e^{2t} + c_2 e^{-3t} \\ 2c_1 e^{2t} - 3c_2 e^{-3t} \end{pmatrix} = c_1 \begin{pmatrix} e^{2t} \\ 2e^{2t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-3t} \\ -3e^{-3t} \end{pmatrix}. \quad (10.10)$$

A sketch of the phase portrait, indicating several representative trajectories, appears in [Figure 10.1](#). The solutions with  $c_2 = 0$  go out to  $\infty$  along the two rays in the directions  $(1, 2)^T$  and  $(-1, -2)^T$ , whereas those with  $c_1 = 0$  come in to the origin along the rays in the directions  $(1, -3)^T$  and  $(-1, 3)^T$ . All other non-equilibrium solutions move along hyperbolic trajectories whose asymptotes, in forward and backward time, are one of these four rays.

With some practice, one learns to understand the temporal behavior of the solution by studying its phase plane trajectory. We will investigate the qualitative and quantitative behavior of phase plane systems in depth in [Section 10.3](#).

## Exercises

10.1.1. Choose one or more of the following differential equations, and then: (a) Solve the equation directly. (b) Write down its phase plane equivalent, and the general solution to the phase plane system. (c) Plot at least four representative trajectories to illustrate the phase portrait. (d) Choose two trajectories in your phase portrait and graph the corresponding solution curves  $u(t)$ . Explain in your own words how the orbit and the solution graph are related. (i)  $\ddot{u} + 4u = 0$ , (ii)  $\ddot{u} - 4u = 0$ , (iii)  $\ddot{u} + 2\dot{u} + u = 0$ , (iv)  $\ddot{u} + 4\dot{u} + 3u = 0$ , (v)  $\ddot{u} - 2\dot{u} + 10u = 0$ .

10.1.2. (a) Convert the third order equation  $\frac{d^3u}{dt^3} + 3\frac{d^2u}{dt^2} + 4\frac{du}{dt} + 12u = 0$  into a first order system in three variables by setting  $u_1 = u$ ,  $u_2 = \dot{u}$ ,  $u_3 = \ddot{u}$ . (b) Solve the equation directly, and then use this to write down the general solution to your first order system. (c) What is the dimension of the solution space?

10.1.3. Convert the second order coupled system of ordinary differential equations

$$\ddot{u} = a\dot{u} + b\dot{v} + cu + dv, \quad \ddot{v} = p\dot{u} + q\dot{v} + ru + sv,$$

into a first order system involving four variables.

- ◇ 10.1.4. (a) Show that if  $\mathbf{u}(t)$  solves  $\dot{\mathbf{u}} = A\mathbf{u}$ , then its *time reversal*, defined as  $\mathbf{v}(t) = \mathbf{u}(-t)$ , solves  $\dot{\mathbf{v}} = B\mathbf{v}$ , where  $B = -A$ . (b) Explain why the two systems have the same phase portraits, but the direction of motion along the trajectories is reversed. (c) Apply time reversal to the system(s) you derived in [Exercise 10.1.1](#). (d) What is the effect of time reversal on the original second order equation?
- ◇ 10.1.5. A first order linear system  $\dot{u} = au + bv$ ,  $\dot{v} = cu + dv$ , can be converted into a single second order differential equation by the following device. Assuming that  $b \neq 0$ , solve the system for  $v$  and  $\dot{v}$  in terms of  $u$  and  $\dot{u}$ . Then differentiate your equation for  $v$  with respect to  $t$ , and eliminate  $\dot{v}$  from the resulting pair of equations. The result is a second order ordinary differential equation for  $u(t)$ . (a) Write out the second order equation in terms of the coefficients  $a, b, c, d$  of the first order system. (b) Show that there is a one-to-one correspondence between solutions of the system and solutions of the scalar differential equation. (c) Use this method to solve the following linear systems, and sketch the resulting phase portraits. (i)  $\dot{u} = v$ ,  $\dot{v} = -u$ , (ii)  $\dot{u} = 2u + 5v$ ,  $\dot{v} = -u$ , (iii)  $\dot{u} = 4u - v$ ,  $\dot{v} = 6u - 3v$ , (iv)  $\dot{u} = u + v$ ,  $\dot{v} = u - v$ , (v)  $\dot{u} = v$ ,  $\dot{v} = 0$ . (d) Show how to obtain a second order equation satisfied by  $v(t)$  by an analogous device. Are the second order equations for  $u$  and for  $v$  the same? (e) Discuss how you might proceed if  $b = 0$ .

10.1.6. (a) Show that if  $\mathbf{u}(t)$  solves  $\dot{\mathbf{u}} = A\mathbf{u}$ , then  $\mathbf{v}(t) = \mathbf{u}(2t)$  solves  $\dot{\mathbf{v}} = B\mathbf{v}$ , where  $B = 2A$  .

(b) How are the solution trajectories of the two systems related?

◇ 10.1.7. Let  $A$  be a constant  $n \times n$  matrix. Let  $\mathbf{u}(t)$  be a solution to the system  $\frac{d\mathbf{u}}{dt} = A\mathbf{u}$ .

(a) Show that its derivatives  $\frac{d^k \mathbf{u}}{dt^k}$  for  $k = 1, 2, \dots$ , are also solutions.

(b) Show that  $\frac{d^k \mathbf{u}}{dt^k} = A^k \mathbf{u}$ .

10.1.8. *True or false:* Each solution to a phase plane system moves at a constant speed along its trajectory.

10.1.9. *True or false:* The phase plane trajectories (10.10) for  $(c_1, c_2)^T \neq \mathbf{0}$  are hyperbolas.

♣ 10.1.10. Use a three-dimensional graphics package to plot solution curves  $(t, u_1(t), u_2(t))^T$  of the phase plane systems in Exercise 10.1.1. Discuss their shape and explain how they are related to the phase plane trajectories.

## Existence and Uniqueness

Before delving further into our subject, it will help to briefly summarize the basic existence and uniqueness theorems as they apply to linear systems of ordinary differential equations. Even though we will study only the constant coefficient case in detail, these results are equally applicable to non-autonomous systems, and so — but only in this subsection — we allow the coefficient matrix to depend continuously on  $t$ . A key fact is that a system of  $n$  first order ordinary differential equations requires  $n$  initial conditions — one for each variable — in order to uniquely specify its solution. More precisely:

**Theorem 10.3.** Let  $A(t)$  be an  $n \times n$  matrix and  $\mathbf{f}(t)$  an  $n$ -component column vector each of whose entries is a continuous functions on the interval<sup>†</sup>  $a < t < b$ . Set an initial time  $a < t_0 < b$  and an initial vector  $\mathbf{b} \in \mathbb{R}^n$ . Then the *initial value problem*

$$\frac{d\mathbf{u}}{dt} = A(t)\mathbf{u} + \mathbf{f}(t), \quad \mathbf{u}(t_0) = \mathbf{b}, \quad (10.11)$$

admits a unique solution  $\mathbf{u}(t)$  that is defined for all  $a < t < b$ .

For completeness, we have included an inhomogeneous forcing term  $\mathbf{f}(t)$  in the system. We will not prove Theorem 10.3, which is a direct consequence of the more general existence and uniqueness theorem for nonlinear systems of ordinary differential equations. Full details can be found in most texts on ordinary differential equations, including [7, 22, 36]. In the homogeneous case, when  $\mathbf{f}(t) \equiv \mathbf{0}$ , uniqueness of solutions implies that the solution with zero initial conditions,  $\mathbf{u}(t_0) = \mathbf{0}$ , is the trivial zero solution:  $\mathbf{u}(t) \equiv \mathbf{0}$  for all  $t$ . In other words, if you start at an equilibrium, you remain there for all time. Moreover, you can never arrive at equilibrium in a finite amount of time, since if  $\mathbf{u}(t_1) = \mathbf{0}$ , then, again by uniqueness,  $\mathbf{u}(t) \equiv \mathbf{0}$  for all  $t < t_1$  (and  $\geq t_1$ , too).

Uniqueness has another important consequence: linear independence of solutions needs be checked only at a single point.

<sup>†</sup> We allow  $a$  and  $b$  to be infinite.

**Lemma 10.4.** The solutions  $\mathbf{u}_1(t), \dots, \mathbf{u}_k(t)$  to a first order homogeneous linear system  $\dot{\mathbf{u}} = A(t)\mathbf{u}$  are linearly independent if and only if their initial values  $\mathbf{u}_1(t_0), \dots, \mathbf{u}_k(t_0)$  are linearly independent vectors in  $\mathbb{R}^n$ .

*Proof:* If the solutions were linearly dependent, one could find (constant) scalars  $c_1, \dots, c_k$ , not all zero, such that

$$\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \dots + c_k \mathbf{u}_k(t) \equiv \mathbf{0}. \quad (10.12)$$

The equation holds, in particular, at  $t = t_0$ ,

$$\mathbf{u}(t_0) = c_1 \mathbf{u}_1(t_0) + \dots + c_k \mathbf{u}_k(t_0) = \mathbf{0}. \quad (10.13)$$

This immediately proves linear dependence of the initial vectors.

Conversely, if the initial values  $\mathbf{u}_1(t_0), \dots, \mathbf{u}_k(t_0)$  are linearly dependent, then (10.13) holds for some  $c_1, \dots, c_k$ , not all zero. Linear superposition implies that the self-same linear combination  $\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \dots + c_k \mathbf{u}_k(t)$  is a solution to the system, with zero initial condition. By uniqueness,  $\mathbf{u}(t) \equiv \mathbf{0}$  for all  $t$ , and so (10.12) holds, proving linear dependence of the solutions. *Q.E.D.*

**Warning.** This result is *not* true if the functions are not solutions to a *first order* linear system. For example,  $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$ ,  $\mathbf{u}_2(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$ , are linearly independent vector-valued functions, but, at time  $t = 0$ , the vectors  $\mathbf{u}_1(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{u}_2(0)$  are linearly dependent. Even worse,  $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$ ,  $\mathbf{u}_2(t) = \begin{pmatrix} t \\ t^2 \end{pmatrix}$ , define linearly dependent vectors at every specified value of  $t$ . Nevertheless, as vector-valued functions, they are linearly independent. (Why?) In view of Lemma 10.4, neither pair of vector-valued functions can be solutions to a common first order homogeneous linear system.

The next result tells us how many different solutions are required in order to construct the general solution by linear superposition.

**Theorem 10.5.** Let  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$  be  $n$  linearly independent solutions to the homogeneous system of  $n$  first order linear ordinary differential equations  $\dot{\mathbf{u}} = A(t)\mathbf{u}$ . Then the general solution is a linear combination  $\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \dots + c_n \mathbf{u}_n(t)$  depending on  $n$  arbitrary constants  $c_1, \dots, c_n$ .

*Proof:* If we have  $n$  linearly independent solutions  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$ , then Lemma 10.4 implies that, at the initial time  $t_0$ , the vectors  $\mathbf{u}_1(t_0), \dots, \mathbf{u}_n(t_0)$  are linearly independent, and hence form a basis for  $\mathbb{R}^n$ . This means that we can express an arbitrary initial condition

$$\mathbf{u}(t_0) = \mathbf{b} = c_1 \mathbf{u}_1(t_0) + \dots + c_n \mathbf{u}_n(t_0)$$

as a linear combination of the initial vectors. Superposition and uniqueness of solutions implies that the corresponding solution to the initial value problem (10.11) is given by the same linear combination

$$\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \dots + c_n \mathbf{u}_n(t).$$

We conclude that every solution to the ordinary differential equation can be written in the prescribed form, which thus forms the general solution. *Q.E.D.*

## Complete Systems

Thus, given a system of  $n$  homogeneous linear differential equations  $\dot{\mathbf{u}} = A\mathbf{u}$ , the immediate goal is to find  $n$  linearly independent solutions. Each eigenvalue  $\lambda$  and eigenvector  $\mathbf{v}$  of its (constant) coefficient matrix  $A$  leads to an exponential eigensolution  $\mathbf{u}(t) = e^{\lambda t} \mathbf{v}$ . The eigensolutions will be linearly independent if and only if the eigenvectors are — this follows directly from Lemma 10.4. Thus, if the  $n \times n$  matrix admits an eigenvector basis, i.e., it is complete, then we have the requisite number of solutions, and hence have solved the differential equation.

**Theorem 10.6.** If the  $n \times n$  matrix  $A$  is complete, then the general (complex) solution to the autonomous linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  is given by

$$\mathbf{u}(t) = c_1 e^{\lambda_1 t} \mathbf{v}_1 + \cdots + c_n e^{\lambda_n t} \mathbf{v}_n, \quad (10.14)$$

where  $\mathbf{v}_1, \dots, \mathbf{v}_n$  are the eigenvector basis,  $\lambda_1, \dots, \lambda_n$  the corresponding eigenvalues. The constants  $c_1, \dots, c_n$  are uniquely specified by the initial conditions  $\mathbf{u}(t_0) = \mathbf{b}$ .

*Proof:* Since the eigenvectors are linearly independent, the eigensolutions define linearly independent vectors  $\mathbf{u}_1(0) = \mathbf{v}_1, \dots, \mathbf{u}_n(0) = \mathbf{v}_n$  at the initial time  $t = 0$ . Lemma 10.4 implies that the eigensolutions  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$  are, indeed, linearly independent. Hence, we know  $n$  linearly independent solutions, and the result is an immediate consequence of Theorem 10.5. *Q.E.D.*

**Example 10.7.** Let us solve the initial value problem

$$\begin{aligned} \dot{u}_1 &= -2u_1 + u_2, & u_1(0) &= 3, \\ \dot{u}_2 &= 2u_1 - 3u_2, & u_2(0) &= 0. \end{aligned}$$

The coefficient matrix is  $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$ . A straightforward computation produces its eigenvalues and eigenvectors:

$$\lambda_1 = -4, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \lambda_2 = -1, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Theorem 10.6 assures us that the corresponding eigensolutions

$$\mathbf{u}_1(t) = e^{-4t} \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad \mathbf{u}_2(t) = e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

form a basis for the two-dimensional solution space. The general solution is an arbitrary linear combination

$$\mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = c_1 e^{-4t} \begin{pmatrix} 1 \\ -2 \end{pmatrix} + c_2 e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1 e^{-4t} + c_2 e^{-t} \\ -2c_1 e^{-4t} + c_2 e^{-t} \end{pmatrix},$$

where  $c_1, c_2$  are constant scalars. Once we have the general solution in hand, the final step is to determine the values of  $c_1, c_2$  in order to satisfy the initial conditions. Evaluating the solution at  $t = 0$ , we find that we need to solve the linear system

$$u_1(0) = c_1 + c_2 = 3, \quad u_2(0) = -2c_1 + c_2 = 0,$$

for  $c_1 = 1, c_2 = 2$ . Thus, the (unique) solution to the initial value problem is

$$u_1(t) = e^{-4t} + 2e^{-t}, \quad u_2(t) = -2e^{-4t} + 2e^{-t}. \quad (10.15)$$

Note that both components of the solution decay exponentially fast to 0 as  $t \rightarrow \infty$ .

**Example 10.8.** Consider the linear initial value problem

$$\begin{aligned}\dot{u}_1 &= u_1 + 2u_2, & u_1(0) &= 2, \\ \dot{u}_2 &= u_2 - 2u_3, & u_2(0) &= -1, \\ \dot{u}_3 &= 2u_1 + 2u_2 - u_3. & u_3(0) &= -2.\end{aligned}$$

The coefficient matrix is  $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -2 \\ 2 & 2 & -1 \end{pmatrix}$ . In Example 8.9, we computed its eigenvalues and eigenvectors:

$$\begin{aligned}\lambda_1 &= -1, & \lambda_2 &= 1 + 2i, & \lambda_3 &= 1 - 2i, \\ \mathbf{v}_1 &= \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 1 \\ -i \\ 1 \end{pmatrix}.\end{aligned}$$

The corresponding eigensolutions to the system are

$$\mathbf{u}_1(t) = e^{-t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \quad \hat{\mathbf{u}}_2(t) = e^{(1+2i)t} \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix}, \quad \hat{\mathbf{u}}_3(t) = e^{(1-2i)t} \begin{pmatrix} 1 \\ -i \\ 1 \end{pmatrix}.$$

The first solution is real, but the second and third, while perfectly valid solutions, are complex-valued, and hence not as convenient to work with if, as in most applications, we are ultimately after real functions. But, since the underlying linear system is real, the general reality principle of Theorem 7.48 tells us that a complex solution can be broken up into its real and imaginary parts, each of which is a *real* solution. Here, applying Euler's formula (3.92) to the complex exponential, we obtain

$$\hat{\mathbf{u}}_2(t) = e^{(1+2i)t} \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix} = (e^t \cos 2t + i e^t \sin 2t) \begin{pmatrix} 1 \\ i \\ 1 \end{pmatrix} = \begin{pmatrix} e^t \cos 2t \\ -e^t \sin 2t \\ e^t \cos 2t \end{pmatrix} + i \begin{pmatrix} e^t \sin 2t \\ e^t \cos 2t \\ e^t \sin 2t \end{pmatrix}.$$

The final two vector-valued functions are independent real solutions, as you can readily check. In this manner, we have produced three linearly independent real solutions

$$\mathbf{u}_1(t) = \begin{pmatrix} -e^{-t} \\ e^{-t} \\ e^{-t} \end{pmatrix}, \quad \mathbf{u}_2(t) = \begin{pmatrix} e^t \cos 2t \\ -e^t \sin 2t \\ e^t \cos 2t \end{pmatrix}, \quad \mathbf{u}_3(t) = \begin{pmatrix} e^t \sin 2t \\ e^t \cos 2t \\ e^t \sin 2t \end{pmatrix},$$

which, by Theorem 10.5, form a basis for the three-dimensional solution space to our system. The general solution can be written as a linear combination:

$$\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + c_2 \mathbf{u}_2(t) + c_3 \mathbf{u}_3(t) = \begin{pmatrix} -c_1 e^{-t} + c_2 e^t \cos 2t + c_3 e^t \sin 2t \\ c_1 e^{-t} - c_2 e^t \sin 2t + c_3 e^t \cos 2t \\ c_1 e^{-t} + c_2 e^t \cos 2t + c_3 e^t \sin 2t \end{pmatrix}.$$

The constants  $c_1, c_2, c_3$  are uniquely prescribed by imposing initial conditions. In our case, the solution satisfying

$$\mathbf{u}(0) = \begin{pmatrix} -c_1 + c_2 \\ c_1 + c_3 \\ c_1 + c_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix} \quad \text{results in} \quad \begin{aligned}c_1 &= -2, \\ c_2 &= 0, \\ c_3 &= 1.\end{aligned}$$

Thus, the solution to the original initial value problem is

$$\mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix} = \begin{pmatrix} 2e^{-t} + e^t \sin 2t \\ -2e^{-t} + e^t \cos 2t \\ -2e^{-t} + e^t \sin 2t \end{pmatrix}.$$

Incidentally, the third complex eigensolution also produces two real solutions, but these reproduce the ones we have already listed, since it is the complex conjugate of the second eigensolution, and so  $\hat{\mathbf{u}}_3(t) = \mathbf{u}_2(t) - i\mathbf{u}_3(t)$ . In general, when solving real systems, you need to deal with only one eigenvalue from each complex conjugate pair to construct a complete system of real solutions.

## Exercises

10.1.11. Find the solution to the system of differential equations  $\frac{du}{dt} = 3u + 4v$ ,  $\frac{dv}{dt} = 4u - 3v$ , with initial conditions  $u(0) = 3$  and  $v(0) = -2$ .

10.1.12. Find the general real solution to the following systems of differential equations:

$$\begin{array}{lll} (a) \begin{aligned} \dot{u}_1 &= u_1 + 9u_2, \\ \dot{u}_2 &= u_1 + 3u_2; \end{aligned} & (b) \begin{aligned} \dot{x}_1 &= 4x_1 + 3x_2, \\ \dot{x}_2 &= 3x_1 - 4x_2; \end{aligned} & (c) \begin{aligned} \dot{y}_1 &= y_1 - y_2, \\ \dot{y}_2 &= 2y_1 + 3y_2; \\ \dot{y}_1 &= y_2, \\ \dot{y}_2 &= 3y_1 + 2y_3, \\ \dot{y}_3 &= -y_2; \end{aligned} & \begin{aligned} \dot{x}_1 &= 3x_1 - 8x_2 + 2x_3, \\ \dot{x}_2 &= -x_1 + 2x_2 + 2x_3, \\ \dot{x}_3 &= x_1 - 4x_2 + 2x_3; \end{aligned} & \begin{aligned} \dot{u}_1 &= u_1 - 3u_2 + 11u_3, \\ \dot{u}_2 &= 2u_1 - 6u_2 + 16u_3, \\ \dot{u}_3 &= u_1 - 3u_2 + 7u_3. \end{aligned} \end{array}$$

10.1.13. Solve the following initial value problems: (a)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ;  
 (b)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$ ;  
 (c)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ;  
 (d)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 3 & -3 \\ 2 & 2 & -7 \\ 0 & 3 & -4 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ ;  
 (e)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 2 & 1 & -6 \\ -1 & 0 & 4 \\ 0 & -1 & -2 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(\pi) = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$ ;  
 (f)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(2) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ ;  
 (g)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 2 & 1 & -1 & 0 \\ -3 & -2 & 0 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 1 & -1 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ -1 \\ 2 \\ 1 \end{pmatrix}$ .

10.1.14. (a) Find the solution to the system  $\frac{dx}{dt} = -x + y$ ,  $\frac{dy}{dt} = -x - y$ , that has initial conditions  $x(0) = 1$ ,  $y(0) = 0$ . (b) Sketch a phase portrait of the system that shows several typical solution trajectories, including the solution you found in part (a). Clearly indicate the direction of increasing  $t$  on your curves.

10.1.15. A planar steady-state fluid flow has velocity vector field  $\mathbf{v} = (2x - 3y, x - y)^T$  at position  $\mathbf{x} = (x, y)^T$ . The corresponding fluid motion is described by the differential equation  $\frac{d\mathbf{x}}{dt} = \mathbf{v}$ . A floating object starts out at the point  $(1, 1)^T$ . Find its position after one time unit.

10.1.16. A steady-state fluid flow has velocity vector field  $\mathbf{v} = (-2y, 2x, z)^T$  at position  $\mathbf{x} = (x, y, z)^T$ . Describe the motion of the fluid particles as governed by the differential equation  $\frac{d\mathbf{x}}{dt} = \mathbf{v}$ .

10.1.17. Solve the initial value problem  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -6 & 1 \\ 1 & -6 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ . Explain how orthogonality can help.

10.1.18. (a) Find the eigenvalues and eigenvectors of  $K = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$ .

(b) Verify that the eigenvectors are mutually orthogonal. (c) Based on part (a), is  $K$  positive definite, positive semi-definite, or indefinite? (d) Solve the initial value problem  $\frac{d\mathbf{u}}{dt} = K \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ , using orthogonality to simplify the computations.

10.1.19. Demonstrate that one can also solve the initial value problem in Example 10.8 by writing the solution as a complex linear combination of the complex eigensolutions, and then using the initial conditions to specify the coefficients.

10.1.20. Determine whether the following vector-valued functions are linearly dependent or linearly independent:

- (a)  $\begin{pmatrix} 1 \\ t \end{pmatrix}$ ,  $\begin{pmatrix} -t \\ 1 \end{pmatrix}$ , (b)  $\begin{pmatrix} 1+t \\ t \end{pmatrix}$ ,  $\begin{pmatrix} 1-t^2 \\ t-t^2 \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 \\ t \end{pmatrix}$ ,  $\begin{pmatrix} t \\ 2 \end{pmatrix}$ ,  $\begin{pmatrix} -t \\ t \end{pmatrix}$ , (d)  $\begin{pmatrix} e^{-t} \\ -e^t \end{pmatrix}$ ,  $\begin{pmatrix} -e^{-t} \\ e^t \end{pmatrix}$ ,  
 (e)  $\begin{pmatrix} e^{2t} \cos 3t \\ -e^{2t} \sin 3t \end{pmatrix}$ ,  $\begin{pmatrix} e^{2t} \sin 3t \\ e^{2t} \cos 3t \end{pmatrix}$ , (f)  $\begin{pmatrix} \cos 3t \\ \sin 3t \end{pmatrix}$ ,  $\begin{pmatrix} \sin 3t \\ \cos 3t \end{pmatrix}$ , (g)  $\begin{pmatrix} 1 \\ t \end{pmatrix}$ ,  $\begin{pmatrix} 0 \\ -2 \end{pmatrix}$ ,  $\begin{pmatrix} 3 \\ 1+3t \end{pmatrix}$ ,  
 (h)  $\begin{pmatrix} e^t \\ -e^t \end{pmatrix}$ ,  $\begin{pmatrix} e^t \\ e^t \end{pmatrix}$ ,  $\begin{pmatrix} -e^t \\ e^t \end{pmatrix}$ , (i)  $\begin{pmatrix} e^t \\ te^t \end{pmatrix}$ ,  $\begin{pmatrix} t^2 e^t \\ e^t \end{pmatrix}$ ,  $\begin{pmatrix} te^t \\ t^2 e^t \end{pmatrix}$ ,  $\begin{pmatrix} e^t \\ e^t \end{pmatrix}$ .

◇ 10.1.21. Let  $A$  be a constant matrix. Suppose  $\mathbf{u}(t)$  solves the initial value problem  $\dot{\mathbf{u}} = A\mathbf{u}$ ,  $\mathbf{u}(0) = \mathbf{b}$ . Prove that the solution to the initial value problem  $\dot{\mathbf{u}} = A\mathbf{u}$ ,  $\mathbf{u}(t_0) = \mathbf{b}$ , is equal to  $\tilde{\mathbf{u}}(t) = \mathbf{u}(t - t_0)$ . How are the solution trajectories related?

10.1.22. Suppose  $\mathbf{u}(t)$  and  $\tilde{\mathbf{u}}(t)$  both solve the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$ . (a) Suppose they have the same value  $\mathbf{u}(t_1) = \tilde{\mathbf{u}}(t_1)$  at any one time  $t_1$ . Show that they are, in fact, the same solution:  $\mathbf{u}(t) = \tilde{\mathbf{u}}(t)$  for all  $t$ . (b) What happens if  $\mathbf{u}(t_1) = \tilde{\mathbf{u}}(t_2)$  for some  $t_1 \neq t_2$ ?

*Hint:* See Exercise 10.1.21.

10.1.23. Prove that the general solution to a linear system  $\dot{\mathbf{u}} = \Lambda \mathbf{u}$  with diagonal coefficient matrix  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  is given by  $\mathbf{u}(t) = (c_1 e^{\lambda_1 t}, \dots, c_n e^{\lambda_n t})^T$ .

10.1.24. Show that if  $\mathbf{u}(t)$  is a solution to  $\dot{\mathbf{u}} = A\mathbf{u}$ , and  $S$  is a constant, nonsingular matrix of the same size as  $A$ , then  $\mathbf{v}(t) = S\mathbf{u}(t)$  solves the linear system  $\dot{\mathbf{v}} = B\mathbf{v}$ , where  $B = SAS^{-1}$  is similar to  $A$ .

◇ 10.1.25. (i) Combine Exercises 10.1.23–24 to show that if  $A = S\Lambda S^{-1}$  is diagonalizable, then the solution to  $\dot{\mathbf{u}} = A\mathbf{u}$  can be written as  $\mathbf{u}(t) = S(c_1 e^{\lambda_1 t}, \dots, c_n e^{\lambda_n t})^T$ , where  $\lambda_1, \dots, \lambda_n$  are its eigenvalues and  $S = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$  is the corresponding matrix of eigenvectors.  
 (ii) Write the general solution to the systems in Exercise 10.1.13 in this form.

## The General Case

Summarizing the preceding subsection, if the coefficient matrix of a homogeneous, autonomous first order linear system is complete, then the eigensolutions form a (complex) basis for the solution space. Assuming the coefficient matrix is real, one obtains a real basis by taking the real and imaginary parts of each complex conjugate pair of solutions. In the incomplete cases, the formulas for the basis solutions are a little more intricate, and

involve polynomials as well as (complex) exponentials. Readers who did not cover Section 8.6 are advised to skip ahead to Section 10.2; only Theorem 10.13, which summarizes the key features, will be used in the sequel.

**Example 10.9.** The simplest incomplete case arises as the phase plane equivalent of a scalar ordinary differential equation whose characteristic equation has a repeated root. For example, to directly solve the second order equation

$$\frac{d^2u}{dt^2} - 2 \frac{du}{dt} + u = 0, \quad (10.16)$$

we substitute the usual exponential ansatz  $u = e^{\lambda t}$ , leading to the characteristic equation

$$\lambda^2 - 2\lambda + 1 = 0.$$

There is only one double root,  $\lambda = 1$ , and hence, up to scalar multiple, only one exponential solution  $u_1(t) = e^t$ . For a scalar ordinary differential equation, the second, “missing” solution is obtained by simply multiplying the first by  $t$ , so that  $u_2(t) = t e^t$ . As a result, the general solution to (10.16) is

$$u(t) = c_1 u_1(t) + c_2 u_2(t) = c_1 e^t + c_2 t e^t.$$

As in (10.8), the equivalent phase plane system is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} \mathbf{u}, \quad \text{where} \quad \mathbf{u}(t) = \begin{pmatrix} u(t) \\ \dot{u}(t) \end{pmatrix}.$$

Note that the coefficient matrix is incomplete — it has  $\lambda = 1$  as a double eigenvalue, but only one independent eigenvector, namely  $\mathbf{v} = (1, 1)^T$ . The two linearly independent solutions to the phase plane system can be constructed from the two solutions to the scalar equation. Thus,

$$\mathbf{u}_1(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}, \quad \mathbf{u}_2(t) = \begin{pmatrix} t e^t \\ t e^t + e^t \end{pmatrix}$$

form a basis for the two-dimensional solution space. The first is an eigensolution, while the second includes an additional polynomial factor. Observe that, in contrast to the scalar case, the second solution  $\mathbf{u}_2$  is *not* obtained from the first by merely multiplying by  $t$ .

In general, the eigenvectors of an incomplete matrix fail to form a basis, and, as noted in Section 8.6, can be extended to a Jordan basis. Thus, the key step is to describe the solutions associated with a Jordan chain, cf. Definition 8.47.

**Lemma 10.10.** Suppose  $\mathbf{w}_1, \dots, \mathbf{w}_k$  form a Jordan chain of length  $k$  for the eigenvalue  $\lambda$  of the matrix  $A$ . Then there are  $k$  linearly independent solutions to the corresponding first order system  $\dot{\mathbf{u}} = A\mathbf{u}$  having the form

$$\begin{aligned} \mathbf{u}_1(t) &= e^{\lambda t} \mathbf{w}_1, & \mathbf{u}_2(t) &= e^{\lambda t} (t \mathbf{w}_1 + \mathbf{w}_2), & \mathbf{u}_3(t) &= e^{\lambda t} \left( \frac{1}{2} t^2 \mathbf{w}_1 + t \mathbf{w}_2 + \mathbf{w}_3 \right), \\ \text{and, in general, } \mathbf{u}_j(t) &= e^{\lambda t} \sum_{i=1}^j \frac{t^{j-i}}{(j-i)!} \mathbf{w}_i, & 1 \leq j \leq k. \end{aligned} \quad (10.17)$$

The proof is by direct substitution of the formulas (10.17) into the differential equation, invoking the defining relations (8.46) of the Jordan chain as needed; details are left to the reader. If  $\lambda$  is a complex eigenvalue, then the Jordan chain solutions (10.17) will involve

complex exponentials. As usual, if  $A$  is a real matrix, they can be split into their real and imaginary parts, which are independent real solutions.

**Example 10.11.** The coefficient matrix of the system

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix} \mathbf{u}$$

is incomplete; it has only 2 linearly independent eigenvectors associated with the eigenvalues 1 and  $-2$ . Using the Jordan basis computed in Example 8.52, we produce the following 5 linearly independent solutions:

$$\begin{aligned} \mathbf{u}_1(t) &= e^t \mathbf{v}_1, & \mathbf{u}_2(t) &= e^t(t\mathbf{v}_1 + \mathbf{v}_2), & \mathbf{u}_3(t) &= e^t\left(\frac{1}{2}t^2\mathbf{v}_1 + t\mathbf{v}_2 + \mathbf{v}_3\right), \\ \mathbf{u}_4(t) &= e^{-2t} \mathbf{v}_4, & \mathbf{u}_5(t) &= e^{-2t}(t\mathbf{v}_4 + \mathbf{v}_5), \end{aligned}$$

or, explicitly,

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ -e^t \\ e^t \end{pmatrix}, \quad \begin{pmatrix} 0 \\ e^t \\ 0 \\ -te^t \\ (-1+t)e^t \end{pmatrix}, \quad \begin{pmatrix} 0 \\ te^t \\ 0 \\ \left(1-\frac{1}{2}t^2\right)e^t \\ \left(-t+\frac{1}{2}t^2\right)e^t \end{pmatrix}, \quad \begin{pmatrix} -e^{-2t} \\ e^{-2t} \\ e^{-2t} \\ -2e^{-2t} \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -(1+t)e^{-2t} \\ te^{-2t} \\ te^{-2t} \\ -2(1+t)e^{-2t} \\ e^{-2t} \end{pmatrix}.$$

The first three solutions are associated with the  $\lambda_1 = 1$  Jordan chain, the last two with the  $\lambda_2 = -2$  chain. The eigensolutions are the pure exponentials  $\mathbf{u}_1(t), \mathbf{u}_4(t)$ . The general solution to the system is an arbitrary linear combination of these five basis solutions.

**Proposition 10.12.** Let  $A$  be an  $n \times n$  matrix. Then the Jordan chain solutions (10.17) constructed from a Jordan basis of  $A$  form a basis for the  $n$ -dimensional solution space for the corresponding linear system  $\dot{\mathbf{u}} = A\mathbf{u}$ .

The proof of linear independence of the Jordan chain solutions follows, via Lemma 10.4, from the linear independence of the Jordan basis vectors, which are their initial values.

Important qualitative features can be readily gleaned from the algebraic structure of the solution formulas (10.17). The following result describes the principal classes of solutions of homogeneous autonomous linear systems of ordinary differential equations.

**Theorem 10.13.** Let  $A$  be a real  $n \times n$  matrix. Every real solution to the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  is a linear combination of  $n$  linearly independent solutions appearing in the following four classes:

- (1) If  $\lambda$  is a complete real eigenvalue of multiplicity  $m$ , then there exist  $m$  linearly independent solutions of the form

$$\mathbf{u}_k(t) = e^{\lambda t} \mathbf{v}_k, \quad k = 1, \dots, m,$$

where  $\mathbf{v}_1, \dots, \mathbf{v}_m$  are linearly independent eigenvectors.

- (2) If  $\lambda_{\pm} = \mu \pm i\nu$  form a pair of complete complex conjugate eigenvalues of multiplicity  $m$ , then there exist  $2m$  linearly independent real solutions of the forms

$$\begin{aligned}\mathbf{u}_k(t) &= e^{\mu t} [\cos(\nu t) \mathbf{x}_k - \sin(\nu t) \mathbf{y}_k], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu t} [\sin(\nu t) \mathbf{x}_k + \cos(\nu t) \mathbf{y}_k],\end{aligned}\quad k = 1, \dots, m,$$

where  $\mathbf{v}_k = \mathbf{x}_k \pm i\mathbf{y}_k$  are the associated complex conjugate eigenvectors.

- (3) If  $\lambda$  is an incomplete real eigenvalue of multiplicity  $m$  and  $r$  is the dimension of the eigenspace  $V_{\lambda}$ , then there exist  $m$  linearly independent solutions of the form

$$\mathbf{u}_k(t) = e^{\lambda t} \mathbf{p}_k(t), \quad k = 1, \dots, m,$$

where  $\mathbf{p}_k(t)$  is a vector of polynomials of degree  $\leq m - r$ .

- (4) If  $\lambda_{\pm} = \mu \pm i\nu$  form a pair of incomplete complex conjugate eigenvalues of multiplicity  $m$ , and  $r$  is the common dimension of the two eigenspaces, then there exist  $2m$  linearly independent real solutions

$$\begin{aligned}\mathbf{u}_k(t) &= e^{\mu t} [\cos(\nu t) \mathbf{p}_k(t) - \sin(\nu t) \mathbf{q}_k(t)], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu t} [\sin(\nu t) \mathbf{p}_k(t) + \cos(\nu t) \mathbf{q}_k(t)],\end{aligned}\quad k = 1, \dots, m,$$

where  $\mathbf{p}_k(t), \mathbf{q}_k(t)$  are vectors of polynomials of degree  $\leq m - r$ , whose detailed structure can be gleaned from Lemma 10.10.

As a result, every real solution to a homogeneous linear system of ordinary differential equations is a vector-valued function whose entries are linear combinations of functions of the particular form  $t^k e^{\mu t} \cos \nu t$  and  $t^k e^{\mu t} \sin \nu t$ , i.e., sums of products of exponentials, trigonometric functions, and polynomials. The exponents  $\mu$  are the real parts of the eigenvalues of the coefficient matrix; the trigonometric frequencies  $\nu$  are the imaginary parts of the eigenvalues; nonconstant polynomials appear only if the matrix is incomplete.

## Exercises

- 10.1.26. Find the general solution to the linear system  $\frac{d\mathbf{u}}{dt} = A\mathbf{u}$  for the following incomplete

coefficient matrices: (a)  $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ , (b)  $\begin{pmatrix} 2 & -1 \\ 9 & -4 \end{pmatrix}$ , (c)  $\begin{pmatrix} -1 & -1 \\ 4 & -5 \end{pmatrix}$ ,  
 (d)  $\begin{pmatrix} 4 & -1 & -3 \\ -2 & 1 & 2 \\ 5 & -1 & -4 \end{pmatrix}$ , (e)  $\begin{pmatrix} -3 & 1 & 0 \\ 1 & -3 & -1 \\ 0 & 1 & -3 \end{pmatrix}$ , (f)  $\begin{pmatrix} 3 & 1 & 1 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$ , (g)  $\begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}$ .

- 10.1.27. Find a first order system of ordinary differential equations that has the indicated vector-valued function as a solution: (a)  $\begin{pmatrix} e^t + e^{2t} \\ 2e^t \end{pmatrix}$ , (b)  $\begin{pmatrix} e^{-t} \cos 3t \\ -3e^{-t} \sin 3t \end{pmatrix}$ , (c)  $\begin{pmatrix} 1 \\ t-1 \end{pmatrix}$ ,  
 (d)  $\begin{pmatrix} \sin 2t - \cos 2t \\ \sin 2t + 3 \cos 2t \end{pmatrix}$ , (e)  $\begin{pmatrix} e^{2t} \\ e^{-3t} \\ e^{2t} - e^{-3t} \end{pmatrix}$ , (f)  $\begin{pmatrix} \sin t \\ \cos t \\ 1 \end{pmatrix}$ , (g)  $\begin{pmatrix} t \\ 1-t^2 \\ 1+t \end{pmatrix}$ , (h)  $\begin{pmatrix} e^t \sin t \\ 2e^t \cos t \\ e^t \sin t \end{pmatrix}$ .

- 10.1.28. Which sets of functions in Exercise 10.1.20 can be solutions to a common first order, homogeneous, constant coefficient linear system of ordinary differential equations? If so, find a system they satisfy; if not, explain why not.

10.1.29. Solve the third order equation  $\frac{d^3u}{dt^3} + 3\frac{d^2u}{dt^2} + 4\frac{du}{dt} + 12u = 0$  by converting it into a first order system. Compare your answer with what you found in Exercise 10.1.2.

10.1.30. Solve the second order coupled system of ordinary differential equations  $\ddot{u} = \dot{u} + u - v$ ,  $\ddot{v} = \dot{v} - u + v$ , by converting it into a first order system involving four variables.

10.1.31. Suppose that  $\mathbf{u}(t) \in \mathbb{R}^n$  is a polynomial solution to the constant coefficient linear system  $\dot{\mathbf{u}} = A\mathbf{u}$ . What is the maximal possible degree of  $\mathbf{u}(t)$ ? What can you say about  $A$  when  $\mathbf{u}(t)$  has maximal degree?

◇ 10.1.32. (a) Under the assumption that  $\mathbf{u}_1, \dots, \mathbf{u}_k$  form a Jordan chain for the coefficient matrix  $A$ , prove that the functions (10.17) are solutions to the system  $\dot{\mathbf{u}} = A\mathbf{u}$ .  
 (b) Prove that they are linearly independent.

## 10.2 Stability of Linear Systems

With the general solution formulas in hand, we are now ready to study the qualitative features of first order linear dynamical systems. Our primary focus will be on stability properties of the equilibrium solution(s). A solution to an autonomous system of first order ordinary differential equations  $\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$  is called an *equilibrium solution* if it remains constant for all  $t$ , so  $\mathbf{u}(t) \equiv \mathbf{u}^*$ . Since its derivative vanishes, this implies that the *equilibrium point*  $\mathbf{u}^*$  satisfies  $\mathbf{f}(\mathbf{u}^*) = \mathbf{0}$ . In particular, for a homogeneous linear system  $\dot{\mathbf{u}} = A\mathbf{u}$ , the origin  $\mathbf{u}^* \equiv \mathbf{0}$  is always an equilibrium point, meaning that a solution that starts out at  $\mathbf{0}$  remains there. The complete set of equilibrium solutions consists of all points  $\mathbf{u}^* \in \ker A$  in the kernel of the coefficient matrix, and so the set of equilibrium solutions forms a subspace — indeed, an invariant subspace — of the configuration space.

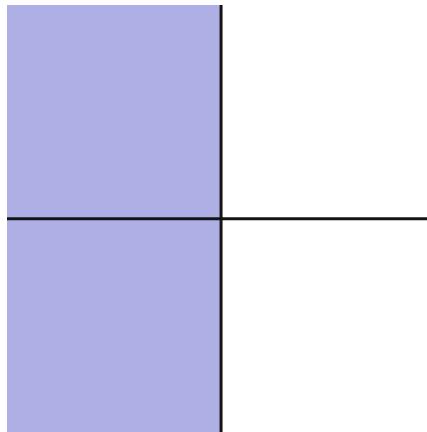
In physical applications, the stability properties of equilibrium solutions is of crucial importance; see the discussion at the beginning of Chapter 5. In general, an equilibrium point is *stable* if *every* solution that starts out nearby stays nearby. An equilibrium is called *asymptotically stable* if the nearby solutions converge to it as time increases. The formal mathematical definitions are as follows.

**Definition 10.14.** An equilibrium solution  $\mathbf{u}^*$  to an autonomous system of first order ordinary differential equations  $\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$  is called

- *stable* if for every sufficiently small  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that every solution  $\mathbf{u}(t)$  having initial conditions within distance  $\delta > \|\mathbf{u}(t_0) - \mathbf{u}^*\|$  of the equilibrium remains within distance  $\varepsilon > \|\mathbf{u}(t) - \mathbf{u}^*\|$  for all  $t \geq t_0$ .
- *asymptotically stable* if it is stable and, in addition, there exists  $\varepsilon_0 > 0$  such that whenever  $\|\mathbf{u}(t_0) - \mathbf{u}^*\| < \varepsilon_0$ , then  $\mathbf{u}(t) \rightarrow \mathbf{u}^*$  as  $t \rightarrow \infty$ .

Thus, although solutions nearby a stable equilibrium point may drift slightly farther away, they must remain relatively close. In the case of asymptotic stability, they will eventually return to equilibrium. An equilibrium point is called *globally stable* if the stability condition holds for *all*  $\varepsilon > 0$ . It is called *globally asymptotically stable* if *every* solution converges to the equilibrium point:  $\mathbf{u}(t) \rightarrow \mathbf{u}^*$  as  $t \rightarrow \infty$ .

In the case of a linear system, local (asymptotic) stability implies global (asymptotic) stability. This is because, by linearity, if  $\mathbf{u}(t)$  is a solution, then so is the scalar multiple  $c\mathbf{u}(t)$  for all  $c \in \mathbb{R}$ , and hence every solution can be scaled to one that remains nearby the



**Figure 10.2.** The Left Half-Plane.

equilibrium point. We will henceforth omit the redundant term “global” when discussing the stability of a linear system. We will also focus our attention on the particular equilibrium solution  $\mathbf{u}^* = \mathbf{0}$ .

**Remark.** The stability and asymptotic stability of an equilibrium solution are independent of the choice of norm in the definition (although this will affect the dependence of  $\delta$  on  $\varepsilon$ ). This follows from the equivalence of norms described in Theorem 3.17.

The starting point is a simple calculus lemma, whose proof is left to the reader.

**Lemma 10.15.** Let  $\mu, \nu$  be real and  $k \geq 0$ . A function of the form

$$f(t) = t^k e^{\mu t} \cos \nu t \quad \text{or} \quad t^k e^{\mu t} \sin \nu t \quad (10.18)$$

will decay to zero for large  $t$ , so  $\lim_{t \rightarrow \infty} f(t) = 0$ , if and only if  $\mu < 0$ . The function remains bounded, so  $|f(t)| \leq C$  for some constant  $C$ , for all  $t \geq 0$  if and only if either  $\mu < 0$ , or  $\mu = 0$  and  $k = 0$ .

Loosely put, exponential decay will always overwhelm polynomial growth, while the trigonometric sine and cosine functions remain neutrally bounded. Now, in the solution to our linear system, the functions (10.18) come from the eigenvalues  $\lambda = \mu + i\nu$  of the coefficient matrix. The lemma implies that the asymptotic behavior of the solutions, and hence the stability of the system, depends on the sign of  $\mu = \operatorname{Re} \lambda$ . If  $\mu < 0$ , then the solutions decay to zero at an exponential rate as  $t \rightarrow \infty$ . If  $\mu > 0$ , then the solutions become unbounded as  $t \rightarrow \infty$ . In the borderline case  $\mu = 0$ , the solutions remain bounded, provided that they don’t involve any powers of  $t$ .

Thus, in order that the equilibrium zero solution be *asymptotically stable*, all the eigenvalues must satisfy  $\mu = \operatorname{Re} \lambda < 0$ . Or, stated another way, all eigenvalues must lie in the *left half-plane* — the subset of the complex plane  $\mathbb{C}$  to the left of the imaginary axis sketched in [Figure 10.2](#). In this manner, we have demonstrated the fundamental asymptotic stability criterion<sup>†</sup> for linear systems.

---

<sup>†</sup> This is *not* the same as the stability criterion for linear iterative systems, which requires that the eigenvalues of the coefficient matrix lie in the inside the unit circle, cf. [Theorem 9.12](#).

**Theorem 10.16.** A first order autonomous homogeneous linear system of ordinary differential equations  $\dot{\mathbf{u}} = A\mathbf{u}$  has an asymptotically stable zero solution if and only if all the eigenvalues  $\lambda$  of its coefficient matrix  $A$  lie in the left half-plane:  $\operatorname{Re} \lambda < 0$ . If  $A$  has one or more eigenvalues with positive real part,  $\operatorname{Re} \lambda > 0$ , then the zero solution is unstable.

**Example 10.17.** Consider the system

$$\frac{du}{dt} = 2u - 6v + w, \quad \frac{dv}{dt} = 3u - 3v - w, \quad \frac{dw}{dt} = 3u - v - 3w.$$

The coefficient matrix  $A = \begin{pmatrix} 2 & -6 & 1 \\ 3 & -3 & -1 \\ 3 & -1 & -3 \end{pmatrix}$  is found to have eigenvalues

$$\lambda_1 = -2, \quad \lambda_2 = -1 + i\sqrt{6}, \quad \lambda_3 = -1 - i\sqrt{6},$$

with respective real parts  $-2, -1, -1$ . The Stability Theorem 10.16 implies that the equilibrium solution  $u_* \equiv v_* \equiv w_* \equiv 0$  is asymptotically stable. Indeed, every solution involves the functions  $e^{-2t}$ ,  $e^{-t} \cos \sqrt{6}t$ , and  $e^{-t} \sin \sqrt{6}t$ , all of which decay to 0 at an exponential rate. The latter two have the slowest decay rate, and so most solutions to the linear system go to  $\mathbf{0}$  in proportion to  $e^{-t}$ , i.e., at an exponential rate determined by the least negative real part.

The final statement is a special case of the following general result, whose proof is left to the reader.

**Proposition 10.18.** If  $\mathbf{u}(t)$  is any solution to  $\dot{\mathbf{u}} = A\mathbf{u}$ , then  $\|\mathbf{u}(t)\| \leq C e^{at}$  for all  $t \geq t_0$  and for all  $a > a^* = \max \{ \operatorname{Re} \lambda \mid \lambda \text{ is an eigenvalue of } A \}$ , where the constant  $C > 0$  depends on the solution and choice of norm. If the eigenvalue(s)  $\lambda$  achieving the maximum,  $\operatorname{Re} \lambda = a^*$ , are complete, then one can set  $a = a^*$ .

Asymptotic stability implies that the solutions return to equilibrium; *stability* only requires them to stay nearby. The appropriate eigenvalue criterion is readily established.

**Theorem 10.19.** A first order linear, homogeneous, constant-coefficient system of ordinary differential equations (10.1) has a stable zero solution if and only if all its eigenvalues satisfy  $\operatorname{Re} \lambda \leq 0$ , and, moreover, any eigenvalue lying on the imaginary axis, so  $\operatorname{Re} \lambda = 0$ , is complete, meaning that it has as many independent eigenvectors as its multiplicity.

*Proof:* The proof is the same as before, based on Theorem 10.13 and the decay properties in Lemma 10.15. All the eigenvalues with negative real part lead to exponentially decaying solutions — even if they are incomplete. If the coefficient matrix has a complete zero eigenvalue, then the corresponding eigensolutions are all constant, and hence trivially bounded. On the other hand, if 0 is an incomplete eigenvalue, then the associated Jordan chain solutions involve non-constant polynomials, and become unbounded as  $t \rightarrow \pm\infty$ . Similarly, if a purely imaginary eigenvalue is complete, then the associated solutions only involve trigonometric functions, and hence remain bounded, whereas the solutions associated with an incomplete purely imaginary eigenvalue contain polynomials in  $t$  multiplying sines and cosines, and hence cannot remain bounded. *Q.E.D.*

A particularly important class of systems consists of the linear *gradient flows*

$$\frac{d\mathbf{u}}{dt} = -K\mathbf{u}, \quad (10.19)$$

in which  $K$  is a symmetric, positive definite matrix. According to Theorem 8.35, all the eigenvalues of  $K$  are real and positive, and so the eigenvalues of the negative definite coefficient matrix  $-K$  for the gradient flow system (10.19) are real and negative. Applying Theorem 10.16, we conclude that the zero solution to any gradient flow system (10.19) with negative definite coefficient matrix  $-K$  is asymptotically stable. If the coefficient matrix is negative semi-definite, the the equilibrium solutions are stable, since the eigenvalues are necessarily complete.

**Example 10.20.** On applying the test we learned in Chapter 3, the matrix  $K = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$  is seen to be positive definite. The associated gradient flow is

$$\frac{du}{dt} = -u - v, \quad \frac{dv}{dt} = -u - 5v. \quad (10.20)$$

The eigenvalues and eigenvectors of  $-K = \begin{pmatrix} -1 & -1 \\ -1 & -5 \end{pmatrix}$  are

$$\lambda_1 = -3 + \sqrt{5}, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix}, \quad \lambda_2 = -3 - \sqrt{5}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix}.$$

Therefore, the general solution to the system is

$$\mathbf{u}(t) = c_1 e^{(-3+\sqrt{5})t} \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix} + c_2 e^{(-3-\sqrt{5})t} \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix},$$

or, in components,

$$\begin{aligned} u(t) &= c_1 e^{(-3+\sqrt{5})t} + c_2 e^{(-3-\sqrt{5})t}, \\ v(t) &= c_1 (2 - \sqrt{5}) e^{(-3+\sqrt{5})t} + c_2 (2 + \sqrt{5}) e^{(-3-\sqrt{5})t}. \end{aligned}$$

All solutions tend to zero as  $t \rightarrow \infty$  at the exponential rate prescribed by the least negative eigenvalue, which is  $-3 + \sqrt{5} \simeq -.7639$ . This confirms the asymptotic stability of the gradient flow.

The reason for the term “gradient flow” is that the vector field  $-K\mathbf{u}$  appearing on the right-hand side of (10.19) is, in fact, the negative of the gradient of the quadratic function

$$q(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T K \mathbf{u} = \frac{1}{2} \sum_{i,j=1}^n k_{ij} u_i u_j, \quad \text{so that} \quad \nabla q(\mathbf{u}) = K\mathbf{u}. \quad (10.21)$$

Thus, we can write (10.19) as

$$\frac{d\mathbf{u}}{dt} = -\nabla q(\mathbf{u}). \quad (10.22)$$

For the particular system (10.20),

$$q(u, v) = \frac{1}{2} (u \ v) \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{2} u^2 + uv + \frac{5}{2} v^2,$$

and so the gradient flow is given by

$$\frac{du}{dt} = -\frac{\partial q}{\partial u} = -u - v, \quad \frac{dv}{dt} = -\frac{\partial q}{\partial v} = -u - 5v.$$

As you learn in multivariable calculus, [2, 78], the gradient  $\nabla q$  of a function  $q$  points in the direction of its steepest increase, while its negative  $-\nabla q$  points in the direction of

steepest decrease. Thus, the solutions to the gradient flow system (10.22) will decrease  $q(\mathbf{u})$  as rapidly as possible, tending to its minimum at  $\mathbf{u}^* = \mathbf{0}$ . For instance, if  $q(u, v)$  represents the height of a hill at position  $(u, v)$ , then the solutions to (10.22) are the paths of steepest descent followed by, say, water flowing down the hill (provided we ignore inertial effects). In physical applications, the quadratic function (10.21) often represents the potential energy in the system, and the gradient flow models the natural behavior of systems that seek to minimize their energy as rapidly as possible.

**Example 10.21.** Another extremely important class of dynamical systems comprises the Hamiltonian systems, first developed by the nineteenth-century Irish mathematician William Rowan Hamilton, who also discovered quaternions, developed in Exercise 7.2.23. In particular, a planar *Hamiltonian system* takes the form

$$\frac{du}{dt} = \frac{\partial H}{\partial v}, \quad \frac{dv}{dt} = -\frac{\partial H}{\partial u}, \quad (10.23)$$

where  $H(u, v)$  is known as the *Hamiltonian function*. If

$$H(u, v) = \frac{1}{2} a u^2 + b u v + \frac{1}{2} c v^2 \quad (10.24)$$

is a quadratic form, then the corresponding Hamiltonian system

$$\dot{u} = b u + c v, \quad \dot{v} = -a u - b v, \quad (10.25)$$

is homogeneous linear, with coefficient matrix  $A = \begin{pmatrix} b & c \\ -a & -b \end{pmatrix}$ . The associated characteristic equation is

$$\det(A - \lambda I) = \lambda^2 + (ac - b^2) = 0.$$

If  $H$  is positive or negative definite, then  $ac - b^2 > 0$ , and so the eigenvalues are purely imaginary:  $\lambda = \pm i\sqrt{ac - b^2}$  and complete, since they are simple. Thus, the stability criterion of Theorem 10.19 holds, and we conclude that planar Hamiltonian systems with a definite Hamiltonian function are stable. On the other hand, if  $H$  is indefinite, then the coefficient matrix has one positive and one negative eigenvalue, and hence the Hamiltonian system is unstable.

In physical applications, the Hamiltonian function  $H(u, v)$  represents the energy of the system. According to Exercise 10.2.22, the Hamiltonian energy function is automatically conserved, meaning that it is constant on every solution:  $H(u(t), v(t)) = \text{constant}$ . This means that the solutions move along its level sets; in the stable cases these are bounded ellipses, whereas in the unstable cases they are unbounded hyperbolas.

**Remark.** The equations of classical mechanics, such as motion of masses (sun, planets, comets, etc.) under gravitational attraction, can all be formulated as Hamiltonian systems, [31]. Moreover, the Hamiltonian formulation is a crucial first step in the physical process of quantizing the classical mechanical equations to determine the quantum mechanical equations of motion, [54].

## Exercises

10.2.1. Classify the following systems according to whether the origin is (i) asymptotically

stable, (ii) stable, or (iii) unstable: (a)  $\frac{du}{dt} = -2u - v, \frac{dv}{dt} = u - 2v$ ; (b)  $\frac{du}{dt} = 2u - 5v, \frac{dv}{dt} = u - v$ ; (c)  $\frac{du}{dt} = -u - 2v, \frac{dv}{dt} = 2u - 5v$ ; (d)  $\frac{du}{dt} = -2v, \frac{dv}{dt} = 8u$ ;

- (e)  $\frac{du}{dt} = -2u - v + w, \frac{dv}{dt} = -u - 2v + w, \frac{dw}{dt} = -3u - 3v + 2w;$   
 (f)  $\frac{du}{dt} = -u - 2v, \frac{dv}{dt} = 6u + 3v - 4w, \frac{dw}{dt} = 4u - 3w;$   
 (g)  $\frac{du}{dt} = 2u - v + 3w, \frac{dv}{dt} = u - v + w, \frac{dw}{dt} = -4u + v - 5w;$   
 (h)  $\frac{du}{dt} = u + v - w, \frac{dv}{dt} = -2u - 3v + 3w, \frac{dw}{dt} = -v + w.$

10.2.2. Write out the formula for the general real solution to the system in Example 10.17 and verify its stability.

10.2.3. Write out and solve the gradient flow system corresponding to the following quadratic forms: (a)  $u^2 + v^2$ , (b)  $uv$ , (c)  $4u^2 - 2uv + v^2$ , (d)  $2u^2 - uv - 2uw + 2v^2 - vw + 2w^2$ .

10.2.4. Write out and solve the Hamiltonian systems corresponding to the first three quadratic forms in Exercise 10.2.3. Which of them are stable?

10.2.5. Which of the following  $2 \times 2$  systems are gradient flows? Which are Hamiltonian systems? In each case, discuss the stability of the zero solution.

- (a)  $\dot{u} = -2u + v, \dot{v} = u - 2v,$  (b)  $\dot{u} = u - 2v, \dot{v} = -2u + v,$  (c)  $\dot{u} = v, \dot{v} = u,$  (d)  $\dot{u} = -v, \dot{v} = u,$  (e)  $\dot{u} = -u - 2v, \dot{v} = -2u - v.$

10.2.6. (a) Show that the matrix  $A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}$  has  $\lambda = \pm i$  as incomplete complex conjugate eigenvalues. (b) Find the general real solution to  $\dot{\mathbf{u}} = A\mathbf{u}$ .

(c) Explain the behavior of a typical solution. Why is the zero solution not stable?

10.2.7. Let  $A$  be a real  $3 \times 3$  matrix, and assume that the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  has a periodic solution of period  $P$ . Prove that every periodic solution of the system has period  $P$ . What other types of solutions can there be? Is the zero solution necessarily stable?

10.2.8. Are the conclusions of Exercise 10.2.7 valid when  $A$  is a  $4 \times 4$  matrix?

10.2.9. Let  $A$  be a real  $5 \times 5$  matrix, and assume that  $A$  has eigenvalues  $i, -i, -2, -1$  (and no others). Is the zero solution to the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  necessarily stable? Explain. Does your answer change if  $A$  is  $6 \times 6$ ?

10.2.10. Prove that if  $A$  is strictly diagonally dominant and each diagonal entry is negative, then the zero equilibrium solution to the linear system of ordinary differential equations  $\dot{\mathbf{u}} = A\mathbf{u}$  is asymptotically stable.

10.2.11. *True or false:* The system  $\dot{\mathbf{u}} = -H_n \mathbf{u}$ , where  $H_n$  is the  $n \times n$  Hilbert matrix (1.72), is asymptotically stable.

10.2.12. *True or false:* If the zero solution of the linear system of differential equations  $\dot{\mathbf{u}} = A\mathbf{u}$  is asymptotically stable, so is the zero solution of the linear iterative system  $\mathbf{u}^{(k+1)} = A\mathbf{u}^{(k)}$  with the same coefficient matrix.

10.2.13. Let  $\mathbf{u}(t)$  solve  $\dot{\mathbf{u}} = A\mathbf{u}$ . Let  $\mathbf{v}(t) = \mathbf{u}(-t)$  be its time reversal.

(a) Write down the linear system  $\dot{\mathbf{v}} = B\mathbf{v}$  satisfied by  $\mathbf{v}(t)$ . Then classify the following statements as *true* or *false*. As always, explain your answers. (b) If  $\dot{\mathbf{u}} = A\mathbf{u}$  is asymptotically stable, then  $\dot{\mathbf{v}} = B\mathbf{v}$  is unstable. (c) If  $\dot{\mathbf{u}} = A\mathbf{u}$  is unstable, then  $\dot{\mathbf{v}} = B\mathbf{v}$  is asymptotically stable. (d) If  $\dot{\mathbf{u}} = A\mathbf{u}$  is stable, then  $\dot{\mathbf{v}} = B\mathbf{v}$  is stable.

10.2.14. *True or false:* (a) If  $\text{tr } A > 0$ , then the system  $\dot{\mathbf{u}} = A\mathbf{u}$  is unstable.

(b) If  $\det A > 0$ , then the system  $\dot{\mathbf{u}} = A\mathbf{u}$  is unstable.

10.2.15. *True or false:* If  $K$  is positive semi-definite, then the zero solution to  $\dot{\mathbf{u}} = -K\mathbf{u}$  is stable.

10.2.16. *True or false:* If  $A$  is a symmetric matrix, then the system  $\dot{\mathbf{u}} = -A^2\mathbf{u}$  has an asymptotically stable equilibrium solution.

10.2.17. Consider the differential equation  $\dot{\mathbf{u}} = -K\mathbf{u}$ , where  $K$  is positive semi-definite.

(a) Find all equilibrium solutions. (b) Prove that all non-constant solutions decay exponentially fast to some equilibrium. What is the decay rate? (c) Is the origin stable, asymptotically stable, or unstable? (d) Prove that, as  $t \rightarrow \infty$ , the solution  $\mathbf{u}(t)$  converges to the orthogonal projection of its initial vector  $\mathbf{a} = \mathbf{u}(0)$  onto  $\ker K$ .

10.2.18. Suppose that  $\mathbf{u}(t)$  satisfies the gradient flow system (10.22).

(a) Prove that  $\frac{d}{dt} q(\mathbf{u}) = -\|K\mathbf{u}\|^2$ .

(b) Explain why if  $\mathbf{u}(t)$  is any nonconstant solution to the gradient flow, then  $q(\mathbf{u}(t))$  is a strictly decreasing function of  $t$ , thus quantifying how fast a gradient flow decreases energy.

10.2.19. Let  $H(u, v) = au^2 + buv + cv^2$  be a quadratic function. (a) Prove that the non-equilibrium trajectories of the associated Hamiltonian system and those of the gradient flow are mutually orthogonal, i.e., they always intersect at right angles. (b) Verify this result for the particular quadratic functions (i)  $u^2 + 3v^2$ , (ii)  $uv$ , by drawing representative trajectories of both systems on the same graph.

10.2.20. *True or false:* If the Hamiltonian system for  $H(u, v)$  is stable, then the corresponding gradient flow  $\dot{\mathbf{u}} = -\nabla H$  is stable.

10.2.21. *True or false:* A nonzero linear  $2 \times 2$  gradient flow cannot be a Hamiltonian flow.

◇ 10.2.22. The law of *conservation of energy* states that the energy in a Hamiltonian system is constant on solutions. (a) Prove that if  $\mathbf{u}(t)$  satisfies the Hamiltonian system (10.23), then  $H(\mathbf{u}(t)) = c$  is a constant, and hence solutions  $\mathbf{u}(t)$  move along the level sets of the Hamiltonian or energy function. Explain how the value of  $c$  is determined by the initial conditions. (b) Plot the level curves of the particular Hamiltonian function  $H(u, v) = u^2 - 2uv + 2v^2$  and verify that they coincide with the solution trajectories.

10.2.23. *True or false:* A nonzero linear  $2 \times 2$  gradient flow cannot be a Hamiltonian system.

10.2.24. (a) Explain how to solve the inhomogeneous system  $\frac{d\mathbf{u}}{dt} = A\mathbf{u} + \mathbf{b}$  when  $\mathbf{b}$  is a constant vector belonging to  $\text{img } A$ . *Hint:* Look at  $\mathbf{v}(t) = \mathbf{u}(t) - \mathbf{u}^*$  where  $\mathbf{u}^*$  is an equilibrium solution. (b) Use your method to solve

$$(i) \frac{du}{dt} = u - 3v + 1, \quad \frac{dv}{dt} = -u - v, \quad (ii) \frac{du}{dt} = 4v + 2, \quad \frac{dv}{dt} = -u - 3.$$

◇ 10.2.25. Prove Lemma 10.15.

◇ 10.2.26. Prove Proposition 10.18.

## 10.3 Two-Dimensional Systems

The two-dimensional case is particularly instructive, since it is relatively easy to analyze, but already manifests most of the key phenomena to be found in higher dimensions. Moreover, the solutions can be easily pictured and their behavior understood through their phase portraits. In this section, we will present a complete classification of the possible qualitative behaviors of real, planar linear dynamical systems.

Setting  $\mathbf{u}(t) = (u(t), v(t))^T$ , a first order planar homogeneous linear system has the explicit form

$$\frac{du}{dt} = au + bv, \quad \frac{dv}{dt} = cu + dv, \quad (10.26)$$

where  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is the (constant) coefficient matrix. As in Section 10.1, we will refer to the  $uv$ -plane as the *phase plane*. In particular, the phase plane equivalents (10.8) of second order scalar equations form a subclass thereof.

According to (8.21), the characteristic equation for the given  $2 \times 2$  matrix is

$$\det(A - \lambda I) = \lambda^2 - \tau \lambda + \delta = 0, \quad (10.27)$$

where

$$\tau = \operatorname{tr} A = a + d, \quad \delta = \det A = ad - bc, \quad (10.28)$$

are, respectively, the trace and the determinant of  $A$ . The eigenvalues, and hence the nature of the solutions, are almost entirely determined by these two quantities. The sign of the *discriminant*

$$\Delta = \tau^2 - 4\delta = (\operatorname{tr} A)^2 - 4 \det A = (a - d)^2 + 4bc \quad (10.29)$$

determines whether the eigenvalues

$$\lambda_{\pm} = \frac{\tau \pm \sqrt{\Delta}}{2} \quad (10.30)$$

are real or complex, and thereby plays a key role in the classification.

Let us summarize the different possibilities as distinguished by their qualitative behavior. Each category will be illustrated by a representative phase portrait, which displays several typical solution trajectories in the phase plane. A complete portrait gallery of planar systems can be found in [Figure 10.3](#).

### Distinct Real Eigenvalues

The coefficient matrix  $A$  has two distinct real eigenvalues  $\lambda_1 < \lambda_2$  if and only if the discriminant is positive:  $\Delta > 0$ . In this case, the solutions take the exponential form

$$\mathbf{u}(t) = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2, \quad (10.31)$$

where  $\mathbf{v}_1, \mathbf{v}_2$  are the eigenvectors and  $c_1, c_2$  are arbitrary constants, to be determined by the initial conditions. Let  $V_k = \{c \mathbf{v}_k \mid c \in \mathbb{R}\}$  for  $k = 1, 2$ , denote the two “eigenlines”, i.e., the one-dimensional eigenspaces.

The asymptotic behavior of the solutions is governed by the eigenvalues. There are five qualitatively different cases, depending upon their signs. These are listed by their descriptive name, followed by the required conditions on the discriminant, trace, and determinant of the coefficient matrix that serve to prescribe the form of the eigenvalues.

Ia. *Stable Node*:  $\Delta > 0, \operatorname{tr} A < 0, \det A > 0$ .

If  $\lambda_1 < \lambda_2 < 0$  are both negative, then  $\mathbf{0}$  is an asymptotically *stable node*. The solutions all tend to  $\mathbf{0}$  as  $t \rightarrow \infty$ . Since the first exponential  $e^{\lambda_1 t}$  decreases much faster than the second  $e^{\lambda_2 t}$ , the first term in the solution (10.31) will soon become negligible, and hence  $\mathbf{u}(t) \approx c_2 e^{\lambda_2 t} \mathbf{v}_2$  when  $t$  is large, provided  $c_2 \neq 0$ . Such solutions will arrive at the origin along curves tangent to the eigenline  $V_2$ , including those with  $c_1 = 0$ , which move directly along the eigenline. On the other hand, the solutions with  $c_2 = 0$  come in to the origin along the eigenline  $V_1$ , at a faster rate. Conversely, as  $t \rightarrow -\infty$ , all solutions become unbounded:  $\|\mathbf{u}(t)\| \rightarrow \infty$ . In this case, the first exponential grows faster than the second, and so  $\mathbf{u}(t) \approx c_1 e^{\lambda_1 t} \mathbf{v}_1$  for  $t \ll 0$ . In other words, as they escape to  $\infty$ , the solution

trajectories become more and more parallel to the eigenline  $V_1$  — except for those with  $c_1 = 0$ , which remain on the eigenline  $V_2$ .

Ib. *Saddle Point*:  $\Delta > 0$ ,  $\det A < 0$ .

If  $\lambda_1 < 0 < \lambda_2$ , then  $\mathbf{0}$  is an unstable *saddle point*. Solutions (10.31) with  $c_2 = 0$  start out on the eigenline  $V_1$  and go in to  $\mathbf{0}$  as  $t \rightarrow \infty$ , while solutions with  $c_1 = 0$  start on  $V_2$  and go to  $\mathbf{0}$  as  $t \rightarrow -\infty$ . All other solutions become unbounded at both large positive and large negative times. As  $t \rightarrow +\infty$ , they asymptotically approach the *unstable eigenline*  $V_2$ , while as  $t \rightarrow -\infty$ , they approach the *stable eigenline*  $V_1$ .

Ic. *Unstable Node*:  $\Delta > 0$ ,  $\text{tr } A > 0$ ,  $\det A > 0$ .

If the eigenvalues  $0 < \lambda_1 < \lambda_2$  are both positive, then  $\mathbf{0}$  is an *unstable node*. The phase portrait is the same as that of a stable node, but the solution trajectories are traversed in the opposite direction. Time reversal  $t \rightarrow -t$  will convert an unstable node into a stable node and vice versa. Thus, in the unstable case, the solutions all tend to the origin as  $t \rightarrow -\infty$  and become unbounded as  $t \rightarrow \infty$ . Except for the eigensolutions, they asymptotically approach  $V_1$  as  $t \rightarrow -\infty$ , and become parallel to  $V_2$  as  $t \rightarrow \infty$ .

Id. *Stable Line*:  $\Delta > 0$ ,  $\text{tr } A < 0$ ,  $\det A = 0$ .

If  $\lambda_1 < \lambda_2 = 0$ , then every point on the eigenline  $V_2$  associated with the zero eigenvalue is a stable equilibrium point. The other solutions move along straight lines parallel to  $V_1$ , asymptotically approaching one of the equilibrium points on  $V_2$  as  $t \rightarrow \infty$ . On the other hand, as  $t \rightarrow -\infty$ , all solutions except those sitting still on the eigenline move off to  $\infty$ .

Ie. *Unstable Line*:  $\Delta > 0$ ,  $\text{tr } A > 0$ ,  $\det A = 0$ .

This is merely the time reversal of a stable line. If  $0 = \lambda_1 < \lambda_2$ , then every point on the eigenline  $V_1$  is an equilibrium. The other solutions moves off to  $\infty$  along straight lines parallel to  $V_2$  as  $t \rightarrow \infty$ , and tend to an equilibrium on  $V_1$  as  $t \rightarrow -\infty$ .

## Complex Conjugate Eigenvalues

The coefficient matrix  $A$  has two complex conjugate eigenvalues

$$\lambda_{\pm} = \mu \pm i\nu, \quad \text{where} \quad \mu = \frac{1}{2}\tau = \frac{1}{2}\text{tr } A, \quad \nu = \sqrt{-\Delta},$$

if and only if its discriminant is negative:  $\Delta < 0$ . In this case, the real solutions can be written in the phase–amplitude form

$$\mathbf{u}(t) = r e^{\mu t} [\cos(\nu t - \sigma) \mathbf{w} - \sin(\nu t - \sigma) \mathbf{z}], \quad (10.32)$$

where  $\mathbf{w} \pm i\mathbf{z}$  are the complex eigenvectors. As noted in Exercise 8.3.12, the real vectors  $\mathbf{w}, \mathbf{z}$  are always linearly independent. The amplitude  $r$  and phase shift  $\sigma$  are uniquely prescribed by the initial conditions. There are three subcases, depending upon the sign of the real part  $\mu$ , or, equivalently, the sign of the trace of  $A$ .

IIa. *Stable Focus*:  $\Delta < 0$ ,  $\text{tr } A < 0$ .

If  $\mu < 0$ , then  $\mathbf{0}$  is an asymptotically *stable focus*. As  $t \rightarrow \infty$ , the solutions all spiral in to the origin at an exponential rate  $e^{\mu t}$  with a common “frequency”  $\nu$  — meaning it takes time  $2\pi/\nu$  for the solution to spiral once around the origin<sup>†</sup>. On the other hand, as

---

<sup>†</sup> But keep in mind that these solutions are not periodic. Thus,  $2\pi/\nu$  is the time interval between successive intersections of the solution and a fixed ray emanating from the origin, e.g., the positive  $x$ -axis.

$t \rightarrow -\infty$ , the solutions spiral off to  $\infty$  at the same exponential rate whilst maintaining their overall frequency.

IIb. *Center*:  $\Delta < 0$ ,  $\text{tr } A = 0$ .

If  $\mu = 0$ , meaning that the eigenvalues  $\lambda_{\pm} = \pm i\nu$  are purely imaginary, then  $\mathbf{0}$  is a *center*. The solutions all move periodically around elliptical orbits, with common frequency  $\nu$  and hence period  $2\pi/\nu$ . In particular, solutions that start out near  $\mathbf{0}$  stay nearby, and hence a center is a stable, but not asymptotically stable, equilibrium.

IIc. *Unstable Focus*:  $\Delta < 0$ ,  $\text{tr } A > 0$ .

If  $\mu > 0$ , then  $\mathbf{0}$  is an *unstable focus*. The phase portrait is the time reversal of a stable focus, with solutions having an unbounded spiral motion as  $t \rightarrow \infty$ , and spiraling in to the origin as  $t \rightarrow -\infty$ , again at an exponential rate  $e^{\mu t}$  with a common “frequency”  $\nu$ .

## Incomplete Double Real Eigenvalue

The coefficient matrix has a double real eigenvalue  $\lambda = \frac{1}{2}\tau = \frac{1}{2}\text{tr } A$  if and only if the discriminant vanishes:  $\Delta = 0$ . The formula for the solutions depends on whether the eigenvalue  $\lambda$  is complete. If  $\lambda$  is an incomplete eigenvalue, admitting only one independent eigenvector  $\mathbf{v}$ , then the solutions are no longer given by simple exponentials. The general solution formula is

$$\mathbf{u}(t) = (c_1 + c_2 t)e^{\lambda t} \mathbf{v} + c_2 e^{\lambda t} \mathbf{w}, \quad (10.33)$$

where  $(A - \lambda \mathbf{I})\mathbf{w} = \mathbf{v}$ , and so  $\mathbf{v}, \mathbf{w}$  form a Jordan chain for the coefficient matrix. We let  $V = \{c\mathbf{v}\}$  denote the eigenline associated with the genuine eigenvector  $\mathbf{v}$ .

IIIa. *Stable Improper Node*:  $\Delta = 0$ ,  $\text{tr } A < 0$ ,  $A \neq \lambda \mathbf{I}$ .

If  $\lambda < 0$  then  $\mathbf{0}$  is an asymptotically *stable improper node*. Since  $te^{\lambda t}$  is larger than  $e^{\lambda t}$  for  $t > 1$ , when  $c_2 \neq 0$ , the solutions  $\mathbf{u}(t) \approx c_2 te^{\lambda t}$  tend to  $\mathbf{0}$  as  $t \rightarrow \infty$  along a curve that is tangent to the eigenline  $V$ , while the eigensolutions with  $c_2 = 0$  move in to the origin along the eigenline. Similarly, as  $t \rightarrow -\infty$ , the solutions go off to  $\infty$  in the opposite direction from their approach, becoming more and more parallel to the same eigenline.

IIIb. *Linear Motion*:  $\Delta = 0$ ,  $\text{tr } A = 0$ ,  $A \neq \lambda \mathbf{I}$ .

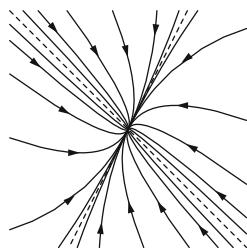
If  $\lambda = 0$ , then every point on the eigenline  $V$  is an unstable equilibrium point. Every other solution is a linear polynomial in  $t$ , and so moves along a straight line parallel to  $V$ , going off to  $\infty$  in either direction.

IIIc. *Unstable Improper Node*:  $\Delta = 0$ ,  $\text{tr } A > 0$ ,  $A \neq \lambda \mathbf{I}$ .

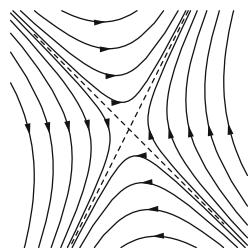
If  $\lambda > 0$ , then  $\mathbf{0}$  is an *unstable improper node*. The phase portrait is the time reversal of the stable improper node. Solutions go off to  $\infty$  as  $t$  increases, becoming progressively more parallel to the eigenline, and tend to the origin tangent to the eigenline as  $t \rightarrow -\infty$ .

## Complete Double Real Eigenvalue

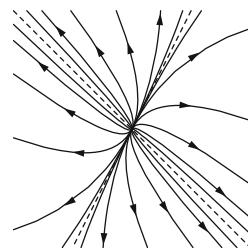
In this case, *every* vector in  $\mathbb{R}^2$  is an eigenvector, and so the real solutions take the form  $\mathbf{u}(t) = e^{\lambda t} \mathbf{v}$ , where  $\mathbf{v}$  is an *arbitrary* constant vector. In fact, this case occurs if and only if  $A = \lambda \mathbf{I}$  is a scalar multiple of the identity matrix.



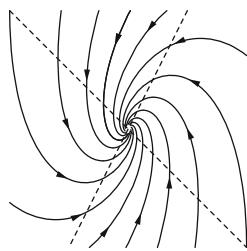
Ia. Stable Node



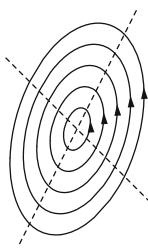
Ib. Saddle Point



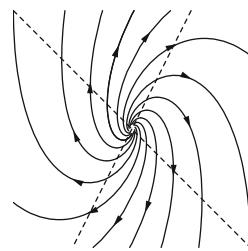
Ic. Unstable Node



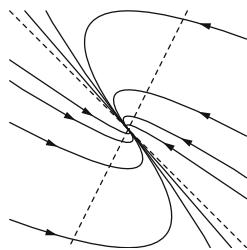
IIa. Stable Focus



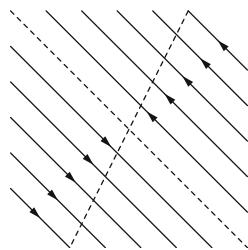
IIb. Center



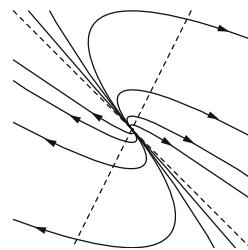
IIc. Unstable Focus



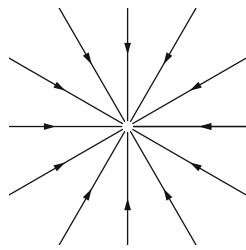
IIIa. Stable Improper Node



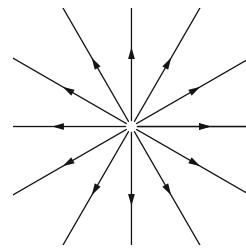
IIIb. Linear Motion



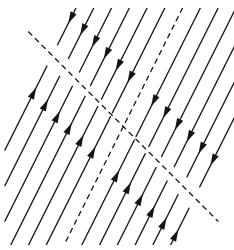
IIIc. Unstable Improper Node



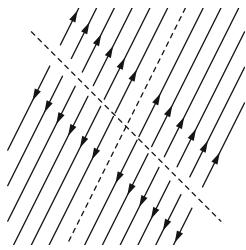
IVa. Stable Star



IVc. Unstable Star

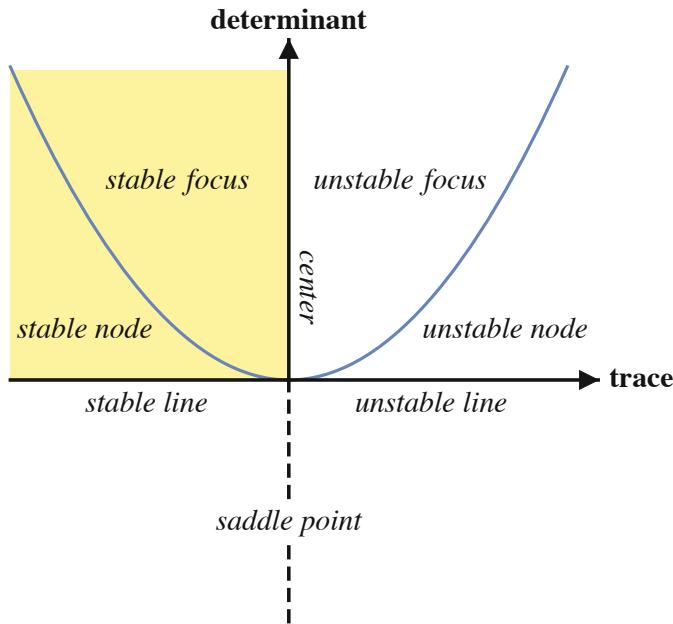


Id. Stable Line



Ie. Unstable Line

**Figure 10.3.** Phase Portraits.



**Figure 10.4.** Stability Regions for Two-Dimensional Linear Systems.

IVa. *Stable Star:*  $A = \lambda I, \quad \lambda < 0.$

If  $\lambda < 0$ , then  $\mathbf{0}$  is an asymptotically *stable star*. The solution trajectories are the rays coming in to the origin, and the solutions go to  $\mathbf{0}$  at a common exponential rate  $e^{\lambda t}$  as  $t \rightarrow \infty$ .

IVb. *Trivial:*  $A = \mathbf{O}.$

If  $\lambda = 0$ , then the only possibility is  $A = \mathbf{O}$ . Now every solution is constant and every point is a (stable) equilibrium point. Nothing happens! This is the only case not pictured in [Figure 10.3](#).

IVc. *Unstable Star:*  $A = \lambda I, \quad \lambda > 0.$

If  $\lambda > 0$ , then  $\mathbf{0}$  is an *unstable star*. The phase portrait is the time reversal of the stable star, and so the solutions move out along rays as  $t \rightarrow \infty$  at an exponential rate  $e^{\lambda t}$ , while tending to  $\mathbf{0}$  as  $t \rightarrow -\infty$ .

[Figure 10.4](#) summarizes the different possibilities, as prescribed by the trace and determinant of the coefficient matrix. The horizontal axis indicates the value of  $\tau = \text{tr } A$ , while the vertical axis refers to  $\delta = \det A$ . Points on the parabola  $\tau^2 = 4\delta$  represent the cases with vanishing discriminant  $\Delta = 0$ , and correspond to either stars or improper nodes — except for the origin, which is either linear motion or trivial. All the asymptotically stable cases lie in the shaded upper left quadrant where  $\text{tr } A < 0$  and  $\det A > 0$ . The borderline points are either stable centers, when  $\text{tr } A = 0, \det A > 0$ , or stable lines, when  $\text{tr } A < 0, \det A = 0$ , or the origin, which may or may not be stable depending upon whether  $A$  is the zero matrix or not. All other values for the trace and determinant result in unstable equilibria. Summarizing:

**Proposition 10.22.** Let  $\tau, \delta$  denote, respectively, the trace and determinant of the coefficient matrix  $A$  of a homogeneous, linear, autonomous planar system of first order ordinary differential equations. Then the system is

- (i) *asymptotically stable* if and only if  $\delta > 0$  and  $\tau < 0$ ;
- (ii) *stable* if and only if  $\delta \geq 0$ ,  $\tau \leq 0$ , and, if  $\delta = \tau = 0$ , also  $A = O$ .

**Remark.** Time reversal  $t \rightarrow -t$  changes the sign of the coefficient matrix  $A \rightarrow -A$ , and hence the sign of its trace,  $\tau \rightarrow -\tau$ , while the determinant  $\delta = \det A = \det(-A)$  is unchanged. Thus, the effect is to reflect Figure 10.4 through the vertical axis, interchanging the stable nodes and spirals with their unstable counterparts, while taking saddle points to saddle points.

In physical applications, the parameters occurring in the dynamical system are usually not known exactly, and so the real dynamics may, in fact, be governed by a slight perturbation of the mathematical model. Thus, it is important to know which systems are *structurally stable*, meaning that their basic qualitative features are preserved under sufficiently small changes in the coefficients. Now, a small perturbation will alter the coefficient matrix slightly, and hence shift its trace and determinant by a comparably small amount. The net effect is to slightly perturb its eigenvalues. Therefore, the question of structural stability reduces to whether the eigenvalues have moved sufficiently far to send the system into a different stability regime. Asymptotically stable systems remain asymptotically stable since a sufficiently small perturbation will not alter the signs of the real parts of its eigenvalues. For a similar reason, unstable systems remain unstable under small perturbations. On the other hand, a borderline stable system — either a center or the trivial system — might become either asymptotically stable or unstable, even under a minuscule perturbation. Such results continue to hold, at least locally, even under suitably small nonlinear perturbations, and thereby lie at the foundations of nonlinear dynamics.

Structural stability requires a bit more, since the overall phase portrait should not significantly change. A system in any of the open regions in the Stability Figure 10.4, i.e., a stable or unstable focus, a stable or unstable node, or a saddle point, is structurally stable, whereas a system that lies on the parabola  $\tau^2 = 4\delta$ , or the horizontal axis, or the positive vertical axis, e.g., an improper node, a stable line, etc., is not, since a small perturbation can easily kick it into either of the adjoining regions. Thus, structural stability requires that the eigenvalues be distinct,  $\lambda_i \neq \lambda_j$ , and have non-zero real part:  $\operatorname{Re} \lambda \neq 0$ . This final result remains valid for linear systems in higher dimensions, [36, 41]. See also [69, 90] and the brief remarks on page 525 concerning the perturbation theory of eigenvalues, in which Wilkinson's spectral condition number quantifies to what extent the eigenvalues are affected by a perturbation of the coefficient matrix.

## Exercises

10.3.1. For each the following: (a) Write the system as  $\dot{\mathbf{u}} = A\mathbf{u}$ . (b) Find the eigenvalues and eigenvectors of  $A$ . (c) Find the general real solution of the system. (d) Draw the phase portrait, indicating its type and stability properties: (i)  $\dot{u}_1 = -u_2$ ,  $\dot{u}_2 = 9u_1$ ,  
(ii)  $\dot{u}_1 = 2u_1 - 3u_2$ ,  $\dot{u}_2 = u_1 - u_2$ , (iii)  $\dot{u}_1 = 3u_1 - 2u_2$ ,  $\dot{u}_2 = 2u_1 - 2u_2$ .

10.3.2. For each of the following systems

$$(i) \quad \dot{\mathbf{u}} = \begin{pmatrix} 2 & -1 \\ 3 & -2 \end{pmatrix} \mathbf{u}, \quad (ii) \quad \dot{\mathbf{u}} = \begin{pmatrix} 1 & -1 \\ 5 & -3 \end{pmatrix} \mathbf{u}, \quad (iii) \quad \dot{\mathbf{u}} = \begin{pmatrix} -3 & 5/2 \\ -5/2 & 2 \end{pmatrix} \mathbf{u}:$$

(a) Find the general real solution. (b) Using the solution formulas obtained in part (a), plot several trajectories of each system. On your graphs, identify the eigenlines (if relevant), and the direction of increasing  $t$  on the trajectories. (c) Write down the type and stability properties of the system.

10.3.3. Classify the following systems, and sketch their phase portraits.

$$(a) \begin{aligned} \frac{du}{dt} &= -u + 4v, & \frac{du}{dt} &= -2u + v, & \frac{du}{dt} &= 5u + 4v, & \frac{du}{dt} &= -3u - 2v, \\ \frac{dv}{dt} &= u - 2v. & \frac{dv}{dt} &= u - 4v. & \frac{dv}{dt} &= u + 2v. & \frac{dv}{dt} &= 3u + 2v. \end{aligned}$$

◇ 10.3.4. Justify the solution formulas (10.32) and (10.33).

10.3.5. Sketch the phase portrait for the following systems: (a)  $\begin{aligned} \dot{u}_1 &= u_1 - 3u_2, \\ \dot{u}_2 &= -3u_1 + u_2. \end{aligned}$

$$(b) \begin{aligned} \dot{u}_1 &= u_1 - 4u_2, & (c) \quad \dot{u}_1 &= u_1 + u_2, & (d) \quad \dot{u}_1 &= u_1 + u_2, & (e) \quad \dot{u}_1 &= \frac{3}{2}u_1 + \frac{5}{2}u_2, \\ \dot{u}_2 &= u_1 - u_2. & \dot{u}_2 &= 4u_1 - 2u_2. & \dot{u}_2 &= u_2. & \dot{u}_2 &= -\frac{5}{2}u_1 + \frac{3}{2}u_2. \end{aligned}$$

10.3.6. Which of the 14 possible two-dimensional phase portraits can occur for the phase plane equivalent (10.8) of a second order scalar ordinary differential equation?

10.3.7. Which of the 14 possible two-dimensional phase portraits can occur

(a) for a linear gradient flow (10.19)? (b) for a linear Hamiltonian system (10.25)?

10.3.8. (a) Solve the initial value problem  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 2 \\ -1 & -3 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ .

(b) Sketch a picture of your solution curve  $\mathbf{u}(t)$ , indicating the direction of motion.

(c) Is the origin (i) stable? (ii) asymptotically stable? (iii) unstable? (iv) none of these? Justify your answer.

## 10.4 Matrix Exponentials

So far, our focus has been on vector-valued solutions  $\mathbf{u}(t)$  to homogeneous linear systems of ordinary differential equations

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}. \quad (10.34)$$

An evident, and, in fact, useful, generalization is to look for *matrix solutions*. Specifically, we seek a matrix-valued function  $U(t)$  that satisfies the corresponding *matrix differential equation*

$$\frac{dU}{dt} = A U(t). \quad (10.35)$$

As with vectors, we compute the derivative of  $U(t)$  by differentiating its individual entries. If  $A$  is an  $n \times n$  matrix, compatibility of matrix multiplication requires that  $U(t)$  be of size  $n \times k$  for some  $k$ . Since matrix multiplication acts column-wise, the individual columns of the matrix solution  $U(t) = (\mathbf{u}_1(t) \dots \mathbf{u}_k(t))$  must solve the original vector system (10.34). Thus, a matrix solution is merely a convenient way of collecting together several different vector solutions. The most important case is that in which  $U(t)$  is a square matrix, of size  $n \times n$ , and so consists of  $n$  vector solutions to the system.

**Example 10.23.** According to Example 10.7, the vector-valued functions

$$\mathbf{u}_1(t) = \begin{pmatrix} e^{-4t} \\ -2e^{-4t} \end{pmatrix}, \quad \mathbf{u}_2(t) = \begin{pmatrix} e^{-t} \\ e^{-t} \end{pmatrix},$$

are both solutions to the linear system

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \mathbf{u}.$$

They can be combined to form the matrix solution

$$U(t) = \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix} \quad \text{satisfying} \quad \frac{dU}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} U.$$

Indeed, by direct calculation

$$\frac{dU}{dt} = \begin{pmatrix} -4e^{-4t} & -e^{-t} \\ 8e^{-4t} & -e^{-t} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} U.$$

The existence and uniqueness theorems are readily adapted to matrix differential equations, and imply that there is a unique matrix solution to the system (10.35) that has initial conditions

$$U(t_0) = B, \quad (10.36)$$

where  $B$  is an  $n \times k$  matrix. Note that the  $j^{\text{th}}$  column  $\mathbf{u}_j(t)$  of the matrix solution  $U(t)$  satisfies the initial value problem

$$\frac{d\mathbf{u}_j}{dt} = A\mathbf{u}_j, \quad \mathbf{u}_j(t_0) = \mathbf{b}_j,$$

where  $\mathbf{b}_j$  denotes the  $j^{\text{th}}$  column of  $B$ .

In the scalar case, the solution to the particular initial value problem

$$\frac{du}{dt} = a u, \quad u(0) = 1,$$

is the ordinary exponential function  $u(t) = e^{ta}$ . Knowing this, we can write down the solution for a more general initial condition

$$u(t_0) = b \quad \text{as} \quad u(t) = b e^{(t-t_0)a}.$$

Let us formulate an analogous initial value problem for a linear system. Recall that, for matrices, the role of the multiplicative unit 1 is played by the identity matrix  $I$ . This inspires the following definition.

**Definition 10.24.** Let  $A$  be a square  $n \times n$  matrix. The *matrix exponential*

$$U(t) = e^{tA} = \exp(tA) \quad (10.37)$$

is the unique  $n \times n$  matrix solution to the initial value problem

$$\frac{dU}{dt} = A U, \quad U(0) = I. \quad (10.38)$$

In particular, one computes  $e^A$  by setting  $t = 1$  in the matrix exponential  $e^{tA}$ . The matrix exponential turns out to enjoy almost all the properties you might expect from its scalar counterpart. First, it is defined for all  $t \in \mathbb{R}$ , and all  $n \times n$  matrices, both real and complex. We can rewrite the defining properties (10.38) in the more suggestive form

$$\frac{d}{dt} e^{tA} = A e^{tA}, \quad e^{0A} = I. \quad (10.39)$$

As in the scalar case, once we know the matrix exponential, we are in a position to solve the general initial value problem.

**Lemma 10.25.** Let  $A$  be an  $n \times n$  matrix. For any  $n \times k$  matrix  $B$ , the solution to the initial value problem

$$\frac{dU}{dt} = AU, \quad U(t_0) = B, \quad \text{is} \quad U(t) = e^{(t-t_0)A} B. \quad (10.40)$$

*Proof:* Since  $B$  is a constant matrix,

$$\frac{dU}{dt} = \frac{d}{dt} [e^{(t-t_0)A} B] = A e^{(t-t_0)A} B = AU,$$

where we applied the chain rule for differentiation and the first property (10.39). Thus,  $U(t)$  is indeed a matrix solution to the system. Moreover, by the second property in (10.39),

$$U(0) = e^{0A} B = I B = B$$

has the correct initial conditions.

*Q.E.D.*

**Remark.** The computation used in the proof is a special instance of the general *Leibniz rule*

$$\frac{d}{dt} [M(t)N(t)] = \frac{dM(t)}{dt} N(t) + M(t) \frac{dN(t)}{dt} \quad (10.41)$$

for the derivative of the product of (compatible) matrix-valued functions  $M(t)$  and  $N(t)$ . The reader is asked to prove this formula in Exercise 10.4.21.

In particular, the solution to the original vector initial value problem

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u}, \quad \mathbf{u}(t_0) = \mathbf{b},$$

can be written in terms of the matrix exponential:

$$\mathbf{u}(t) = e^{(t-t_0)A} \mathbf{b}. \quad (10.42)$$

Thus, the matrix exponential provides us with an alternative formula for the solution of autonomous homogeneous first order linear systems, providing us with valuable new insight.

The next step is to find an algorithm for computing the matrix exponential. The solution formula (10.40) gives a hint. Suppose  $U(t)$  is any  $n \times n$  matrix solution. Then, by uniqueness,  $U(t) = e^{tA} U(0)$ , and hence, provided that  $U(0)$  is a nonsingular matrix,

$$e^{tA} = U(t) U(0)^{-1}, \quad (10.43)$$

since  $e^{0A} = U(0) U(0)^{-1} = I$ , as required. Thus, to construct the exponential of an  $n \times n$  matrix  $A$ , you first need to find a basis of  $n$  linearly independent solutions  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$  to the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  using the eigenvalues and eigenvectors, or, in the incomplete case, the Jordan chains. The resulting  $n \times n$  matrix solution  $U(t) = (\mathbf{u}_1(t) \dots \mathbf{u}_n(t))$  is then used to produce  $e^{tA}$  via formula (10.43).

**Example 10.26.** For the matrix  $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$  considered in Example 10.23, we already constructed the nonsingular matrix solution  $U(t) = \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix}$ . Therefore,

by (10.43), its matrix exponential is

$$e^{tA} = U(t)U(0)^{-1}$$

$$= \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{3}e^{-4t} + \frac{2}{3}e^{-t} & -\frac{1}{3}e^{-4t} + \frac{1}{3}e^{-t} \\ -\frac{2}{3}e^{-4t} + \frac{2}{3}e^{-t} & \frac{2}{3}e^{-4t} + \frac{1}{3}e^{-t} \end{pmatrix}.$$

In particular, we obtain  $e^A = \exp A$  by setting  $t = 1$  in this formula:

$$\exp \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3}e^{-4} + \frac{2}{3}e^{-1} & -\frac{1}{3}e^{-4} + \frac{1}{3}e^{-1} \\ -\frac{2}{3}e^{-4} + \frac{2}{3}e^{-1} & \frac{2}{3}e^{-4} + \frac{1}{3}e^{-1} \end{pmatrix}.$$

Observe that the matrix exponential is *not* obtained by exponentiating the individual matrix entries.

To solve the initial value problem

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \mathbf{u}, \quad \mathbf{u}(0) = \mathbf{b} = \begin{pmatrix} 3 \\ 0 \end{pmatrix},$$

we appeal to formula (10.40), whence

$$\mathbf{u}(t) = e^{tA} \mathbf{b} = \begin{pmatrix} \frac{1}{3}e^{-4t} + \frac{2}{3}e^{-t} & -\frac{1}{3}e^{-4t} + \frac{1}{3}e^{-t} \\ -\frac{2}{3}e^{-4t} + \frac{2}{3}e^{-t} & \frac{2}{3}e^{-4t} + \frac{1}{3}e^{-t} \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{-4t} + 2e^{-t} \\ -2e^{-4t} + 2e^{-t} \end{pmatrix}.$$

This reproduces our earlier solution (10.15).

**Example 10.27.** Suppose  $A = \begin{pmatrix} -1 & -2 \\ 2 & -1 \end{pmatrix}$ . Its characteristic equation

$$\det(A - \lambda I) = \lambda^2 + 2\lambda + 5 = 0 \quad \text{has roots} \quad \lambda = -1 \pm 2i,$$

which are thus the eigenvalues. The corresponding eigenvectors are  $\mathbf{v} = \begin{pmatrix} \pm i \\ 1 \end{pmatrix}$ , leading to the complex conjugate solutions

$$\mathbf{u}_1(t) = \begin{pmatrix} ie^{(-1+2i)t} \\ e^{(-1+2i)t} \end{pmatrix}, \quad \mathbf{u}_2(t) = \begin{pmatrix} -ie^{(-1-2i)t} \\ e^{(-1-2i)t} \end{pmatrix}.$$

We assemble them to form the (complex) matrix solution

$$U(t) = \begin{pmatrix} ie^{(-1+2i)t} & -ie^{(-1-2i)t} \\ e^{(-1+2i)t} & e^{(-1-2i)t} \end{pmatrix}.$$

The corresponding matrix exponential is, therefore,

$$\begin{aligned} e^{tA} &= U(t)U(0)^{-1} = \begin{pmatrix} ie^{(-1+2i)t} & -ie^{(-1-2i)t} \\ e^{(-1+2i)t} & e^{(-1-2i)t} \end{pmatrix} \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{e^{(-1+2i)t} + e^{(-1-2i)t}}{2} & \frac{-e^{(-1+2i)t} + e^{(-1-2i)t}}{2i} \\ \frac{e^{(-1+2i)t} - e^{(-1-2i)t}}{2i} & \frac{e^{(-1+2i)t} + e^{(-1-2i)t}}{2} \end{pmatrix} = \begin{pmatrix} e^{-t} \cos 2t & -e^{-t} \sin 2t \\ e^{-t} \sin 2t & e^{-t} \cos 2t \end{pmatrix}. \end{aligned}$$

Note that the final expression for the matrix exponential is real, as it must be, since  $A$  is a real matrix. (See Exercise 10.4.19.) Also note that it wasn't necessary to find the real solutions to construct the matrix exponential — although this would also have worked and

yielded the same result. Indeed, the two columns of  $e^{tA}$  form a basis for the space of (real) solutions to the linear system  $\dot{\mathbf{u}} = A\mathbf{u}$ .

Let us finish by listing some further important properties of the matrix exponential, all of which are direct analogues of the usual scalar exponential function. Proofs are relegated to the exercises. First, the *multiplicative property* says that

$$e^{(s+t)A} = e^{sA} e^{tA}, \quad \text{for all } s, t \in \mathbb{R}. \quad (10.44)$$

In particular, if we set  $s = -t$ , the left hand side of (10.44) reduces to the identity matrix, in accordance with the second identity in (10.39), and hence

$$e^{-tA} e^{tA} = \mathbf{I}, \quad \text{and hence} \quad e^{-tA} = (e^{tA})^{-1}. \quad (10.45)$$

As a consequence, for any  $A$  and any  $t \in \mathbb{R}$ , the exponential  $e^{tA}$  is a nonsingular matrix.

**Warning.** In general,

$$e^{t(A+B)} \neq e^{tA} e^{tB}. \quad (10.46)$$

Indeed, according to Proposition 10.30, the left- and right-hand sides of (10.46) are equal for all  $t$  if and only if  $AB = BA$  — that is,  $A$  and  $B$  are commuting matrices.

While the matrix exponential can be painful to compute, there is a simple formula for its determinant in terms of the trace of the generating matrix.

**Lemma 10.28.** Let  $A$  be a square matrix. Then  $\det e^{tA} = e^{t \operatorname{tr} A}$ .

*Proof:* According to Exercise 10.4.26, if  $A$  has eigenvalues  $\lambda_1, \dots, \lambda_n$ , then  $e^{tA}$  has eigenvalues  $e^{t\lambda_1}, \dots, e^{t\lambda_n}$ . Moreover, using (8.26), its determinant,  $\det e^{tA}$ , is the product of its eigenvalues, and so

$$\det e^{tA} = e^{t\lambda_1} e^{t\lambda_2} \cdots e^{t\lambda_n} = e^{t(\lambda_1 + \lambda_2 + \cdots + \lambda_n)} = e^{t \operatorname{tr} A},$$

where, by (8.25), we identify the sum of the eigenvalues as the trace of  $A$ . Q.E.D.

For instance, the matrix  $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$  considered above in Example 10.26 has  $\operatorname{tr} A = (-2) + (-3) = -5$ , and hence  $\det e^{tA} = e^{-5t}$ , as you can easily check.

Finally, we note that the standard exponential series is also valid for matrices:

$$e^{tA} = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n = \mathbf{I} + tA + \frac{t^2}{2} A^2 + \frac{t^3}{6} A^3 + \cdots. \quad (10.47)$$

To prove that the series converges, we use the matrix norm convergence criterion in Exercise 9.2.44(c). Indeed, the corresponding series of matrix norms is bounded by the scalar exponential series,

$$\|e^{tA}\| \leq \sum_{n=0}^{\infty} \left\| \frac{t^n}{n!} A^n \right\| = \sum_{n=0}^{\infty} \frac{|t|^n}{n!} \|A^n\| \leq \sum_{n=0}^{\infty} \frac{|t|^n}{n!} \|A\|^n = e^{|t| \|A\|},$$

which converges for all  $t$ , [2, 78], thereby proving convergence. With this in hand, proving that the exponential series satisfies the defining initial value problem (10.39) is straightforward:

$$\frac{d}{dt} \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n = \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!} A^n = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^{n+1} = A \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n.$$

Moreover, at  $t = 0$ , the sum collapses to the identity matrix:  $\mathbf{I} = e^{0A}$ . Thus, formula (10.47) follows from the uniqueness of solutions to the matrix initial value problem.

## Exercises

10.4.1. Find the exponentials  $e^{tA}$  of the following  $2 \times 2$  matrices:

$$(a) \begin{pmatrix} 2 & -1 \\ 4 & -3 \end{pmatrix}, (b) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, (c) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, (d) \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, (e) \begin{pmatrix} -1 & 2 \\ -5 & 5 \end{pmatrix}, (f) \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix}.$$

10.4.2. Determine the matrix exponential  $e^{tA}$  for the following matrices:

$$(a) \begin{pmatrix} 0 & 0 & 0 \\ 2 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, (b) \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}, (c) \begin{pmatrix} -1 & 1 & 1 \\ -2 & -2 & -2 \\ 1 & -1 & -1 \end{pmatrix}, (d) \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

10.4.3. Verify the determinant formula of Lemma 10.28 for the matrices in Exercises 10.4.1 and 10.4.2.

10.4.4. Solve the indicated initial value problems by first exponentiating the coefficient matrix and then applying formula (10.42): (a)  $\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{u}$ ,  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ ,

$$(b) \frac{d\mathbf{u}}{dt} = \begin{pmatrix} 3 & -6 \\ 4 & -7 \end{pmatrix} \mathbf{u}, \quad \mathbf{u}(0) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad (c) \frac{d\mathbf{u}}{dt} = \begin{pmatrix} -9 & -6 & 6 \\ 8 & 5 & -6 \\ -2 & 1 & 3 \end{pmatrix} \mathbf{u}, \quad \mathbf{u}(0) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

10.4.5. Find  $e^A$  when  $A =$

$$(a) \begin{pmatrix} 5 & -2 \\ -2 & 5 \end{pmatrix}, (b) \begin{pmatrix} 1 & -2 \\ 1 & 1 \end{pmatrix}, (c) \begin{pmatrix} 2 & -1 \\ 4 & -2 \end{pmatrix}, (d) \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -5 \end{pmatrix}, (e) \begin{pmatrix} 0 & 1 & -2 \\ -1 & 0 & 2 \\ 2 & -2 & 0 \end{pmatrix}.$$

10.4.6. Let  $A = \begin{pmatrix} 0 & -2\pi \\ 2\pi & 0 \end{pmatrix}$ . Show that  $e^A = \mathbf{I}$ .

10.4.7. What is  $e^{tO}$  where  $O$  is the  $n \times n$  zero matrix?

10.4.8. Find all matrices  $A$  such that  $e^{tA} = O$ .

◇ 10.4.9. Explain in detail why the columns of  $e^{tA}$  form a basis for the solution space to the system  $\dot{\mathbf{u}} = A\mathbf{u}$ .

10.4.10. Let  $A$  be a  $2 \times 2$  matrix such that  $\text{tr } A = 0$  and  $\delta = \sqrt{\det A} > 0$ .

(a) Prove that  $e^A = (\cos \delta) \mathbf{I} + \frac{\sin \delta}{\delta} A$ . Hint: Use Exercise 8.2.52.

(b) Establish a similar formula when  $\det A < 0$ . (c) What if  $\det A = 0$ ?

10.4.11. Show that the origin is an asymptotically stable equilibrium solution to  $\dot{\mathbf{u}} = A\mathbf{u}$  if and only if  $\lim_{t \rightarrow \infty} e^{tA} = O$ .

10.4.12. Let  $A$  be a real square matrix and  $e^A$  its exponential. Under what conditions does the linear system  $\dot{\mathbf{u}} = e^A \mathbf{u}$  have an asymptotically stable equilibrium solution?

10.4.13. True or false: (a)  $e^{A^{-1}} = (e^A)^{-1}$ ; (b)  $e^{A+A^{-1}} = e^A e^{A^{-1}}$ .

◇ 10.4.14. Prove formula (10.44). Hint: Fix  $s$  and prove that, as functions of  $t$ , both sides of the equation define matrix solutions with the same initial conditions. Then use uniqueness.

10.4.15. Prove that  $A$  commutes with its exponential:  $A e^{tA} = e^{tA} A$ .

◇ 10.4.16. (a) Prove that the exponential of the transpose of a matrix is the transpose of its exponential:  $e^{tA^T} = (e^{tA})^T$ . (b) What does this imply about the solutions to the linear systems  $\dot{\mathbf{u}} = A\mathbf{u}$  and  $\dot{\mathbf{v}} = A^T\mathbf{v}$ ?

- ◇ 10.4.17. Prove that if  $A = SBS^{-1}$  are similar matrices, then so are  $e^{tA} = Se^{tB}S^{-1}$ .
- 10.4.18. Prove that  $e^{t(A-\lambda I)} = e^{-t\lambda} e^{tA}$  by showing that both sides are matrix solutions to the same initial value problem.
- ◇ 10.4.19. Let  $A$  be a real matrix. (a) Explain why  $e^A$  is a real matrix. (b) Prove that  $\det e^A > 0$ .
- 10.4.20. Show that  $\text{tr } A = 0$  if and only if  $\det e^{tA} = 1$  for all  $t$ .
- ◇ 10.4.21. Justify the matrix Leibniz rule (10.41) using the formula for matrix multiplication.
- 10.4.22. Prove that if  $U(t)$  is any matrix solution to  $\frac{dU}{dt} = AU$ , then so is  $\tilde{U}(t) = U(t)C$ , where  $C$  is any constant matrix (of compatible size).
- ◇ 10.4.23. Prove that if  $A = \begin{pmatrix} B & \mathbf{0} \\ \mathbf{0} & C \end{pmatrix}$  is a block diagonal matrix, then so is  $e^{tA} = \begin{pmatrix} e^{tB} & \mathbf{0} \\ \mathbf{0} & e^{tC} \end{pmatrix}$ .

- ◇ 10.4.24. (a) Prove that if  $J_{0,n}$  is an  $n \times n$  Jordan block matrix with 0 diagonal entries,

$$\text{cf. (8.49), then } e^{tJ_{0,n}} = \begin{pmatrix} 1 & t & \frac{t^2}{2} & \frac{t^3}{6} & \cdots & \frac{t^n}{n!} \\ 0 & 1 & t & \frac{t^2}{2} & \cdots & \frac{t^{n-1}}{(n-1)!} \\ 0 & 0 & 1 & t & \cdots & \frac{t^{n-2}}{(n-2)!} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

- (b) Determine the exponential of a general Jordan block matrix  $J_{\lambda,n}$ . Hint: Use Exercise 10.4.18. (c) Explain how you can use the Jordan canonical form to compute the exponential of a matrix. Hint: Use Exercise 10.4.23.

- ◇ 10.4.25. Diagonalization provides an alternative method for computing the exponential of a complete matrix. (a) First show that if  $D = \text{diag}(d_1, \dots, d_n)$  is a diagonal matrix, so is  $e^{tD} = \text{diag}(e^{td_1}, \dots, e^{td_n})$ . (b) Second, using Exercise 10.4.17, prove that if  $A = SDS^{-1}$  is diagonalizable, so is  $e^{tA} = Se^{tD}S^{-1}$ . (c) When possible, use diagonalization to compute the exponentials of the matrices in Exercises 10.4.1–2.

- ◇ 10.4.26. (a) Prove that if  $\lambda$  is an eigenvalue of  $A$ , then  $e^{t\lambda}$  is an eigenvalue of  $e^{tA}$ . What is the eigenvector? (b) Show that the eigenvalues have the same multiplicities.
- Hint: Combine the Jordan canonical form (8.51) with Exercises 10.4.24 and 10.4.25.

- ◇ 10.4.27. Let  $A$  be a symmetric matrix with Spectral Decomposition

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_k P_k,$$

- as in (8.37). Prove that

$$e^{tA} = e^{t\lambda_1} P_1 + e^{t\lambda_2} P_2 + \cdots + e^{t\lambda_k} P_k.$$

- ◇ 10.4.28. (a) Show that  $U(t)$  satisfies the matrix differential equation  $\dot{U} = UB$  if and only if  $U(t) = Ce^{tB}$ , where  $C = U(0)$ . (b) If  $U(0)$  is nonsingular, then  $U(t)$  also satisfies a matrix differential equation of the form  $\dot{U} = AU$ . Is  $A = B$ ? Hint: Use Exercise 10.4.17.

- 10.4.29. *True or false:* The solution to the non-autonomous initial value problem

$$\dot{\mathbf{u}} = A(t) \mathbf{u}, \quad \mathbf{u}(0) = \mathbf{b}, \quad \text{is} \quad \mathbf{u}(t) = \exp\left(\int_0^t A(s) ds\right) \mathbf{b}.$$

- ◇ 10.4.30. (a) Suppose  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$  are vector-valued functions whose values at each point  $t$  are linearly independent vectors in  $\mathbb{R}^n$ . Show that they form a basis for the solution space of a homogeneous constant coefficient linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  if and only if each  $d\mathbf{u}_j/dt$  is a linear combination of  $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$ . *Hint:* Use Exercise 10.4.28. (b) Show that a function  $\mathbf{u}(t)$  belongs to the solution space of a homogeneous constant coefficient linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  if and only if  $\frac{d^n \mathbf{u}}{dt^n}$  is a linear combination of  $\mathbf{u}, \frac{d\mathbf{u}}{dt}, \dots, \frac{d^{n-1} \mathbf{u}}{dt^{n-1}}$ . *Hint:* Use Exercise 10.1.7.

- ◇ 10.4.31. By a (natural) *logarithm* of a matrix  $B$  we mean a matrix  $A$  such that  $e^A = B$ .

- (a) Explain why only nonsingular matrices can have a logarithm.  
 (b) Comparing Exercises 10.4.6–7, explain why the matrix logarithm is not unique.

- (c) Find all real logarithms of the  $2 \times 2$  identity matrix  $\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ .  
*Hint:* Use Exercise 10.4.26.

## Applications in Geometry

Matrix exponentials are an effective tool for understanding the linear transformations that appear in geometry and group theory, [93], quantum mechanics, [54], computer graphics and animation, [5, 12, 72], computer vision, [73], and the symmetry analysis of differential equations, [13, 60]. We will only be able to scratch the surface of this important and active area of contemporary mathematical research.

Let  $A$  be an  $n \times n$  matrix. For each  $t \in \mathbb{R}$ , the corresponding exponential  $e^{tA}$  is itself an  $n \times n$  matrix and thus defines a linear transformation on the vector space  $\mathbb{R}^n$ :

$$L_t[\mathbf{x}] = e^{tA} \mathbf{x} \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^n.$$

In this manner, each square matrix  $A$  generates a family of invertible linear transformations, parameterized by  $t \in \mathbb{R}$ . The resulting linear transformations are not arbitrary, but are subject the following three rules:

$$L_t \circ L_s = L_{t+s} = L_s \circ L_t, \quad L_0 = \mathbf{I}, \quad L_{-t} = L_t^{-1}. \quad (10.48)$$

These are merely restatements of three of the basic matrix exponential properties listed in (10.39, 44, 45). In particular, every transformation in the family commutes with every other one.

In geometry, the family of transformations  $L_t = e^{tA}$  is said to form a *one-parameter group*<sup>†</sup>, [60], with  $t$  the parameter, and the matrix  $A$  is referred to as its *infinitesimal generator*. Indeed, by the series formula (10.39) for the matrix exponential,

$$L_t[\mathbf{x}] = e^{tA} \mathbf{x} = \left( \mathbf{I} + tA + \frac{1}{2}t^2A^2 + \dots \right) \mathbf{x} = \mathbf{x} + tA\mathbf{x} + \frac{1}{2}t^2A^2\mathbf{x} + \dots. \quad (10.49)$$

When  $t$  is small, we can truncate the exponential series and approximate the transformation by the linear function

$$F_t[\mathbf{x}] = (\mathbf{I} + tA)\mathbf{x} = \mathbf{x} + tA\mathbf{x} \quad (10.50)$$

defined by the infinitesimal generator. We already made use of such approximations when we discussed the rigid motions and mechanisms of structures in Chapter 6. As  $t$  varies, the

<sup>†</sup> See also Exercise 4.3.24 for the general definition of a group.

group transformations (10.49) typically move a point  $\mathbf{x}$  along a curved trajectory. Under the first order approximation (10.50), the point  $\mathbf{x}$  moves along a straight line in the direction  $\mathbf{b} = A\mathbf{x}$  — the tangent line to the curved trajectory. Thus, *the infinitesimal generator of a one-parameter group prescribes the tangent line approximation to the nonlinear motion prescribed by the group transformations.*

Most of the linear transformations of interest in the above-mentioned applications arise in this fashion. Let's look briefly at a few basic examples.

- (a) When  $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ , then  $e^{tA} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$  represents a shearing transformation. The group laws (10.48) imply that the composition of a shear of magnitude  $s$  and a shear of magnitude  $t$  in the same direction is another shear of magnitude  $s+t$ .
- (b) When  $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ , then  $e^{tA} = \begin{pmatrix} e^t & 0 \\ 0 & e^t \end{pmatrix}$  represents a uniform scaling transformation. Composition and inverses of such scaling transformations are also scalings.
- (c) When  $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ , then  $e^{tA} = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$ , which, for  $t > 0$ , represents a stretch in the  $x$  direction and a contraction in the  $y$  direction.
- (d) When  $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ , then  $e^{tA} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$  is the matrix for a plane rotation, around the origin, by angle  $t$ . The group laws (10.48) say that the composition of a rotation through angle  $s$  followed by a rotation through angle  $t$  is a rotation through angle  $s+t$ , as previously noted in Example 7.12. Also, the inverse of a rotation through angle  $t$  is a rotation through angle  $-t$ .

Observe that the infinitesimal generator of this one-parameter group of plane rotations is a  $2 \times 2$  skew-symmetric matrix. This turns out to be a general fact: rotations in higher dimensions are also generated by skew-symmetric matrices.

**Lemma 10.29.** If  $A^T = -A$  is a skew-symmetric matrix, then, for all  $t \in \mathbb{R}$ , its matrix exponential  $Q(t) = e^{tA}$  is a proper orthogonal matrix.

*Proof:* According to equation (10.45) and Exercise 10.4.16,

$$Q(t)^{-1} = e^{-tA} = e^{tA^T} = (e^{tA})^T = Q(t)^T,$$

which proves orthogonality. Properness,  $\det Q = +1$ , follows from Lemma 10.28 using the fact that  $\text{tr } A = 0$ , since all the diagonal entries of a skew-symmetric matrix are 0. *Q.E.D.*

With some more work, it can be shown that every proper orthogonal matrix is the exponential of some skew-symmetric matrix, albeit not a unique one. Thus, the  $\frac{1}{2}n(n-1)$ -dimensional vector space of  $n \times n$  skew-symmetric matrices generates the group of rotations in  $n$ -dimensional Euclidean space. In the three-dimensional case, the three matrices  $A_x, A_y, A_z$  listed below form a basis and serve to generate, respectively, the one-parameter groups of counterclockwise rotations around the  $x$ -,  $y$ -, and  $z$ -axes:

$$\begin{aligned}
A_x &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, & e^{tA_x} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}, \\
A_y &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, & e^{tA_y} &= \begin{pmatrix} \cos t & 0 & \sin t \\ 0 & 1 & 0 \\ -\sin t & 0 & \cos t \end{pmatrix}, \\
A_z &= \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & e^{tA_z} &= \begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{pmatrix}.
\end{aligned} \tag{10.51}$$

Since every other skew-symmetric matrix can be expressed as a linear combination of  $A_x$ ,  $A_y$ , and  $A_z$ , every rotation can, in a sense, be generated by these three basic types. This reconfirms our earlier observations concerning the number of rigid motions (rotations and translations) experienced by an unattached structure; see Section 6.3 for details.

In the three-dimensional case, it can be shown that every non-zero skew-symmetric  $3 \times 3$  matrix  $A$  is singular, with one-dimensional kernel. Let  $\mathbf{0} \neq \mathbf{v} \in \ker A$  be the null eigenvector. Then the matrix exponentials  $e^{tA}$  form the one-parameter group of rotations around the axis defined by  $\mathbf{v}$ . For instance, referring to (10.51),  $\ker A_x$  is spanned by  $\mathbf{e}_1 = (1, 0, 0)^T$ , reconfirming that it generates the rotations around the  $x$ -axis. Details can be found in Exercise 10.4.38.

Noncommutativity of linear transformations is reflected in the noncommutativity of their infinitesimal generators. Recall, (1.12), that the *commutator* of two  $n \times n$  matrices  $A, B$  is

$$[A, B] = AB - BA. \tag{10.52}$$

Thus,  $A$  and  $B$  commute if and only if  $[A, B] = \mathbf{0}$ . We use the exponential series (10.47) to evaluate the commutator of the corresponding matrix exponentials:

$$\begin{aligned}
[e^{tA}, e^{tB}] &= e^{tA} e^{tB} - e^{tB} e^{tA} \\
&= (I + tA + \frac{1}{2}t^2A^2 + \dots)(I + tB + \frac{1}{2}t^2B^2 + \dots) - \\
&\quad - (I + tB + \frac{1}{2}t^2B^2 + \dots)(I + tA + \frac{1}{2}t^2A^2 + \dots) \\
&= t^2(AB - BA) + \dots = t^2[A, B] + \dots.
\end{aligned} \tag{10.53}$$

In particular, if the groups commute, then  $[A, B] = \mathbf{0}$ . The converse is also true, since if  $AB = BA$  then all terms in the two series commute, and hence the matrix exponentials also commute.

**Proposition 10.30.** The matrix exponentials  $e^{tA}$  and  $e^{tB}$  commute for all  $t$  if and only if the matrices  $A$  and  $B$  commute:

$$e^{tA} e^{tB} = e^{tB} e^{tA} = e^{t(A+B)} \quad \text{provided} \quad AB = BA. \tag{10.54}$$

In particular, the non-commutativity of three-dimensional rotations follows from the non-commutativity of their infinitesimal skew-symmetric generators. For instance, the commutator of the generators of rotations around the  $x$ - and  $y$ -axes is the generator of rotations around the  $z$ -axis:  $[A_x, A_y] = A_z$ , since

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

Hence, to a first order (or, more correctly, second order) approximation, the difference between  $x$  and  $y$  rotations is, interestingly, a  $z$  rotation.

## Exercises

10.4.32. Find the one-parameter groups generated by the following matrices and interpret geometrically: What are the trajectories? What are the fixed points?

$$(a) \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 3 \\ -3 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 0 & -1 \\ 4 & 0 \end{pmatrix}, \quad (e) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

10.4.33. Write down the one-parameter groups generated by the following matrices and interpret. What are the trajectories? What are the fixed points?

$$(a) \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & 0 & -2 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (e) \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

10.4.34. (a) Find the one-parameter group of rotations generated by the skew-symmetric matrix  $A = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$ . (b) As noted above,  $e^{tA}$  represents a family of rotations around a fixed axis in  $\mathbb{R}^3$ . What is the axis?

10.4.35. Choose two of the groups in Exercise 10.4.32 or 10.4.33, and determine whether or not they commute by looking at their infinitesimal generators. Then verify your conclusion by directly computing the commutator of the corresponding matrix exponentials.

- 10.4.36. (a) Prove that the commutator of two upper triangular matrices is upper triangular.  
 (b) Prove that the commutator of two skew-symmetric matrices is skew symmetric.  
 (c) Is the commutator of two symmetric-matrices symmetric?

◇ 10.4.37. Prove that the *Jacobi identity*

$$[[A, B], C] + [[C, A], B] + [[B, C], A] = 0 \quad (10.55)$$

is valid for three  $n \times n$  matrices  $A, B, C$ .

◇ 10.4.38. Let  $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^3$ . (a) Show that the cross product  $L_{\mathbf{v}}[\mathbf{x}] = \mathbf{v} \times \mathbf{x}$  defines a linear transformation on  $\mathbb{R}^3$ . (b) Find the  $3 \times 3$  matrix representative  $A_{\mathbf{v}}$  of  $L_{\mathbf{v}}$  and show that it is skew-symmetric. (c) Show that every non-zero skew-symmetric  $3 \times 3$  matrix defines such a cross product map. (d) Show that  $\ker A_{\mathbf{v}}$  is spanned by  $\mathbf{v}$ .  
 (e) Justify the fact that the matrix exponentials  $e^{tA_{\mathbf{v}}}$  are rotations around the axis  $\mathbf{v}$ . Thus, the cross product with a vector serves as the infinitesimal generator of the one-parameter group of rotations around  $\mathbf{v}$ .

◇ 10.4.39. Given a unit vector  $\|\mathbf{u}\| = 1$  in  $\mathbb{R}^3$ , let  $A = A_{\mathbf{u}}$  be the corresponding skew-symmetric  $3 \times 3$  matrix that satisfies  $A\mathbf{x} = \mathbf{u} \times \mathbf{x}$ , as in Exercise 10.4.38. (a) Prove the *Euler–Rodrigues formula*  $e^{tA} = I + (\sin t)A + (1 - \cos t)A^2$ . Hint: Use the matrix exponential series (10.47).  
 (b) Show that  $e^{tA} = I$  if and only if  $t$  is an integer multiple of  $2\pi$ . (c) Generalize parts (a) and (b) to a non-unit vector  $\mathbf{v} \neq \mathbf{0}$ .

◇ 10.4.40. Let  $A = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ ,  $\mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ . (a) Show that the solution to the linear system

$\dot{\mathbf{x}} = A\mathbf{x}$  represents a rotation of  $\mathbb{R}^3$  around the  $z$ -axis. What is the trajectory of a point  $\mathbf{x}_0$ ? (b) Show that the solution to the inhomogeneous system  $\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}$  represents a screw motion of  $\mathbb{R}^3$  around the  $z$ -axis. What is the trajectory of a point  $\mathbf{x}_0$ ? (c) More generally, given  $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^3$ , show that the solution to  $\dot{\mathbf{x}} = \mathbf{a} \times \mathbf{x} + \mathbf{a}$  represents a family of screw motions along the axis  $\mathbf{a}$ .

10.4.41. Let  $A$  be an  $n \times n$  matrix whose last row has all zero entries. Prove that the last row of  $e^{tA}$  is  $\mathbf{e}_n^T = (0, \dots, 0, 1)$ .

10.4.42. Let  $A = \begin{pmatrix} B & \mathbf{c} \\ \mathbf{0} & 0 \end{pmatrix}$  be in block form, where  $B$  is an  $n \times n$  matrix,  $\mathbf{c} \in \mathbb{R}^n$ , while  $\mathbf{0}$

denotes the zero row vector with  $n$  entries. Show that its matrix exponential is also in

block form  $e^{tA} = \begin{pmatrix} e^{tB} & \mathbf{f}(t) \\ \mathbf{0} & 1 \end{pmatrix}$ . Can you find a formula for  $\mathbf{f}(t)$ ?

◇ 10.4.43. According to Exercise 7.3.10, an  $(n+1) \times (n+1)$  matrix of the block form  $\begin{pmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{pmatrix}$  in which  $A$  is an  $n \times n$  matrix and  $\mathbf{b} \in \mathbb{R}^n$  can be identified with the affine transformation  $F[\mathbf{x}] = A\mathbf{x} + \mathbf{b}$  on  $\mathbb{R}^n$ . Exercise 10.4.42 shows that every matrix in the one-parameter group  $e^{tB}$  generated by  $B = \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0} & 0 \end{pmatrix}$  has such a form, and hence we can identify  $e^{tB}$  as a family of affine maps on  $\mathbb{R}^n$ . Describe the affine transformations of  $\mathbb{R}^2$  generated by the following matrices:

$$(a) \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (b) \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (c) \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad (d) \begin{pmatrix} 1 & 0 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{pmatrix}.$$

## Invariant Subspaces and Linear Dynamical Systems

Invariant subspaces, as per Definition 8.27, play an important role in the study of homogeneous linear dynamical systems. In general, a subset  $S \subset \mathbb{R}^n$  is called *invariant* for the homogeneous linear dynamical system  $\dot{\mathbf{u}} = A\mathbf{u}$  if, whenever the initial condition  $\mathbf{u}(t_0) = \mathbf{b} \in S$ , then the solution  $\mathbf{u}(t) \in S$  for all  $t \in \mathbb{R}$ .

**Proposition 10.31.** If  $V \subset \mathbb{R}^n$  is an invariant subspace of the matrix  $A$ , then it is invariant under the corresponding homogeneous linear dynamical system.

*Proof:* Given that  $\mathbf{b} \in V$ , we have  $A\mathbf{b} \in V$ ,  $A^2\mathbf{b} \in V$ , and, in general,  $A^n\mathbf{b} \in V$  for each  $n \geq 0$ . Thus every term in the matrix exponential series for the solution (10.42), namely

$$\mathbf{u}(t) = e^{(t-t_0)A} \mathbf{b} = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n \mathbf{b},$$

belongs to  $V$  and hence, because  $V$  is closed, so does their sum:  $\mathbf{u}(t) \in V$ . Q.E.D.

As we know, the (complex) invariant subspaces of a complete matrix are spanned by its (complex) eigenvectors. According to the general Stability Theorem 10.13, these come in three flavors, depending upon whether the real part of the corresponding eigenvalue is less than, equal to, or greater than 0. The first kind, with  $\operatorname{Re} \lambda < 0$ , correspond to the asymptotically stable eigensolutions  $\mathbf{u}(t) = e^{\lambda t} \mathbf{v} \rightarrow \mathbf{0}$  as  $t \rightarrow \infty$ . The second kind, with zero real part, correspond to stable eigensolutions that remain bounded for all  $t$ , by completeness. The third kind, with  $\operatorname{Re} \lambda > 0$ , correspond to unstable eigensolutions that become unbounded at an exponential rate as  $t \rightarrow \infty$ . A similar result holds for the corresponding real solutions of a complete real matrix. If the matrix is incomplete, then the solutions corresponding to Jordan chains with eigenvalues having negative real part are also asymptotically stable; those corresponding to Jordan chains with eigenvalues having positive real part remain exponentially unstable. If any purely imaginary eigenvalue is incomplete, then the polynomial factor in front of the corresponding Jordan chain solution

makes it unstable, becoming unbounded at a polynomial rate. An example of the latter behavior is provided by a planar system that has 0 as its incomplete eigenvalue, producing unstable linear motion. The minimum dimension of a (real) system possessing a non-zero, incomplete purely imaginary eigenvalue is 4.

This motivates dissecting the underlying vector space into three invariant subspaces, having only the zero vector in common, that capture the three possible modes of behavior. We state the definition in the real case, leaving the simpler complex version to the reader.

**Definition 10.32.** Let  $A$  be a real  $n \times n$  matrix. We define the following invariant subspaces spanned by the real and imaginary parts of the eigenvectors and Jordan chains corresponding to the eigenvalues with the following properties:

- (i) negative real part: the *stable subspace*  $S \subset \mathbb{R}^n$ ;
- (ii) zero real part: the *center subspace*  $C \subset \mathbb{R}^n$ ;
- (iii) positive real part: the *unstable subspace*  $U \subset \mathbb{R}^n$ .

If there are no eigenvalues of the specified type, the corresponding invariant subspace is trivial. For example, if the associated linear system has asymptotically stable zero solution, then  $S = \mathbb{R}^n$  while  $C = U = \{\mathbf{0}\}$ . The stable, unstable, and center subspaces are complementary, as in Exercise 2.2.24, in the sense that their pairwise intersections are trivial:  $S \cap C = S \cap U = C \cap U = \{\mathbf{0}\}$ , and their sum  $S + C + U = \mathbb{R}^n$ , in the sense that every  $\mathbf{v} \in \mathbb{R}^n$  can be, in fact uniquely, written as a sum  $\mathbf{v} = \mathbf{s} + \mathbf{c} + \mathbf{u}$  of vectors in each subspace:  $\mathbf{s} \in S$ ,  $\mathbf{c} \in C$ ,  $\mathbf{u} \in U$ .

Since each of these subspaces is invariant, if the initial condition belongs to one of them, so does the corresponding solution. In view of the solution formulas in Theorem 10.13, we deduce the following more intrinsic characterizations, in terms of the asymptotic behavior of the solutions to the homogeneous linear dynamical system.

**Theorem 10.33.** Let  $A$  be an  $n \times n$  matrix. Let  $\mathbf{0} \neq \mathbf{b} \in \mathbb{R}^n$ , and let  $\mathbf{u}(t)$  be a solution to the associated initial value problem  $\dot{\mathbf{u}} = A\mathbf{u}$ ,  $\mathbf{u}(t_0) = \mathbf{b}$ . Then  $\mathbf{b}$  and hence  $\mathbf{u}(t)$  are in:

- (i) the stable subspace  $S$  if and only if  $\mathbf{u}(t) \rightarrow \mathbf{0}$  as  $t \rightarrow \infty$ , or, alternatively,  $\|\mathbf{u}(t)\| \rightarrow \infty$  at an exponential rate as  $t \rightarrow -\infty$ ;
- (ii) the center subspace  $C$  if and only if  $\mathbf{u}(t)$  is bounded or  $\|\mathbf{u}(t)\| \rightarrow \infty$  at a polynomial rate as  $t \rightarrow \pm\infty$ ;
- (iii) the unstable subspace  $U$  if and only if  $\|\mathbf{u}(t)\| \rightarrow \infty$  at an exponential rate as  $t \rightarrow \infty$ , or, alternatively,  $\mathbf{u}(t) \rightarrow \mathbf{0}$  as  $t \rightarrow -\infty$ .

**Example 10.34.** For example, the matrix  $A = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & -2 \end{pmatrix}$  has eigenvalues and eigenvectors

$$\lambda_1 = 0, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \lambda_2 = -2, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \quad \lambda_3 = -3, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Thus, the stable subspace is the plane spanned by  $\mathbf{v}_2$  and  $\mathbf{v}_3$ , whose nonzero solutions tend to the origin as  $t \rightarrow \infty$  at an exponential rate; the center subspace is the line spanned by  $\mathbf{v}_1$ , all of whose solutions are constant; the unstable subspace is trivial:  $U = \{\mathbf{0}\}$ . So the origin is a stable, but not asymptotically stable, equilibrium point.

The Center Manifold Theorem, a celebrated result in nonlinear dynamics, [34], states that the above formulated linear splitting into stable, center, and unstable regimes carries

over to nonlinear systems in a neighborhood of an equilibrium point. Roughly speaking, suppose that  $\mathbf{u}_0$  is an equilibrium point of the nonlinear systems of ordinary differential equations  $\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u})$ , so that  $\mathbf{f}(\mathbf{u}_0) = \mathbf{0}$ . Let  $A$  be the *linearization* of  $\mathbf{f}(\mathbf{u})$  at  $\mathbf{u}_0$ , meaning its Jacobian matrix. so  $A = \mathbf{f}'(\mathbf{u}_0) = (\partial f_i / \partial u_j)$ , evaluated at the equilibrium point. Then, in a neighborhood of  $\mathbf{u}_0$ , the dynamical system admits three invariant curved manifolds, meaning curves, surfaces, and their higher-dimensional counterparts, called the *stable*, *center*, and *unstable manifolds* that are tangent to (or equivalently, approximated by) the corresponding invariant subspaces of its linearization matrix  $A$ . Solutions evolving on the stable and unstable manifolds exhibit behaviors similar to those of the linear system. In particular, solutions on the stable manifold converge to the equilibrium,  $\mathbf{u}(t_0) \rightarrow \mathbf{u}_0$  as  $t \rightarrow \infty$ , at an exponential rate governed by the corresponding eigenvalues of  $A$ , while those on the unstable manifold move away from the equilibrium point  $\mathbf{u}_0$  — although one cannot say what happens to them once they exit the neighborhood, once the nonlinear effects take over. Solutions on the center manifold have more subtle dynamical behavior, that depends on the detailed structure of the nonlinear terms. In this manner, one can effectively argue that, near a fixed point, all the interesting dynamics takes place on the center manifold.

---

## Exercises

10.4.44. (a) Given a homogeneous linear dynamical system with invariant stable, unstable, and center subspaces  $S, U, C$ , explain why the origin is asymptotically stable if and only if  $C = U = \{\mathbf{0}\}$ . (b) Is the origin stable if  $U = \{\mathbf{0}\}$  but  $C \neq \{\mathbf{0}\}$ ?

10.4.45. Find the (real) stable, unstable, and center subspaces of the following linear systems:

$$\begin{array}{llll}
 \text{(a)} \quad \dot{u}_1 = 9u_2, & \text{(b)} \quad \dot{x}_1 = 4x_1 + x_2, & \text{(c)} \quad \dot{y}_1 = y_1 - y_2, & \text{(d)} \quad \dot{z}_1 = z_2, \\
 \dot{u}_2 = -u_1; & \dot{x}_2 = 3x_1; & \dot{y}_2 = 2y_1 + 3y_2; & \dot{z}_2 = 3z_1 + 2z_3, \\
 & & & \dot{y}_3 = -z_2; \\
 \text{(e)} \quad \dot{u}_1 = u_1 - 3u_2 + 11u_3, & \text{(f)} \quad \frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 3 & -3 \\ 2 & 2 & -7 \\ 0 & 3 & -4 \end{pmatrix} \mathbf{u}, & \text{(g)} \quad \frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix} \mathbf{u}.
 \end{array}$$

◇ 10.4.46. State and prove a counterpart to Definition 10.32 and Theorem 10.33 for a homogeneous linear iterative system.

---

## Inhomogeneous Linear Systems

We now direct our attention to inhomogeneous linear systems of ordinary differential equations. For simplicity, we consider only first order<sup>†</sup> systems of the form

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u} + \mathbf{f}(t), \tag{10.56}$$

in which  $A$  is a constant  $n \times n$  matrix and  $\mathbf{f}(t)$  is a vector-valued function of  $t$  that can be interpreted as a collection of time-varying external forces acting on the system. According to Theorem 7.38, the solution to the inhomogeneous system will have the general form

$$\mathbf{u}(t) = \mathbf{u}^*(t) + \mathbf{z}(t)$$

---

<sup>†</sup> Higher order systems can, as in the phase plane construction, (10.8), always be converted into first order systems involving additional variables.

where  $\mathbf{u}^*(t)$  is a particular solution, representing a response to the forcing, while  $\mathbf{z}(t)$  is a solution to the corresponding homogeneous system  $\dot{\mathbf{z}} = A\mathbf{z}$ , representing the system's internal motion. Since we now know how to find the solution  $\mathbf{z}(t)$  to the homogeneous system, the only task is to find one particular solution to the inhomogeneous system.

In your first course on ordinary differential equations, you probably encountered a method known as *variation of parameters* for constructing particular solutions of inhomogeneous scalar ordinary differential equations, [7]. The method can be readily adapted to first order systems. Recall that, in the scalar case, to solve the inhomogeneous equation

$$\frac{du}{dt} = au + f(t), \quad \text{we set} \quad u(t) = e^{ta} v(t), \quad (10.57)$$

where the function  $v(t)$  is to be determined. Differentiating, we obtain

$$\frac{du}{dt} = a e^{ta} v(t) + e^{ta} \frac{dv}{dt} = au + e^{ta} \frac{dv}{dt}.$$

Therefore,  $u(t)$  satisfies the differential equation (10.57) if and only if

$$\frac{dv}{dt} = e^{-ta} f(t).$$

Since the right-hand side of the latter is known,  $v(t)$  can be immediately found by a direct integration.

The method can be extended to the vector-valued situation as follows. We replace the scalar exponential by the exponential of the coefficient matrix, setting

$$\mathbf{u}(t) = e^{tA} \mathbf{v}(t), \quad (10.58)$$

where  $\mathbf{v}(t)$  is a vector-valued function that is to be determined. Combining the product rule for matrix multiplication (10.41) with (10.39) yields

$$\frac{d\mathbf{u}}{dt} = \frac{d}{dt} (e^{tA} \mathbf{v}) = \frac{de^{tA}}{dt} \mathbf{v} + e^{tA} \frac{d\mathbf{v}}{dt} = A e^{tA} \mathbf{v} + e^{tA} \frac{d\mathbf{v}}{dt} = A \mathbf{u} + e^{tA} \frac{d\mathbf{v}}{dt}.$$

Comparing with the differential equation (10.56), we conclude that

$$\frac{d\mathbf{v}}{dt} = e^{-tA} \mathbf{f}(t).$$

Integrating<sup>†</sup> both sides from the initial time  $t_0$  to time  $t$  produces, by the Fundamental Theorem of Calculus,

$$\mathbf{v}(t) = \mathbf{v}(t_0) + \int_{t_0}^t e^{-sA} \mathbf{f}(s) ds, \quad \text{where} \quad \mathbf{v}(t_0) = e^{-t_0A} \mathbf{u}(t_0). \quad (10.59)$$

Substituting back into (10.58) leads to a general formula for the solution to the inhomogeneous linear system.

**Theorem 10.35.** The solution to the initial value problem

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u} + \mathbf{f}(t), \quad \mathbf{u}(t_0) = \mathbf{b}, \quad \text{is} \quad \mathbf{u}(t) = e^{(t-t_0)A} \mathbf{b} + \int_{t_0}^t e^{(t-s)A} \mathbf{f}(s) ds. \quad (10.60)$$

---

<sup>†</sup> As with differentiation, vector-valued and matrix-valued functions are integrated entry-wise.

In the solution formula, the integral term can be viewed as a particular solution  $\mathbf{u}^*(t)$ , namely the one satisfying the initial condition  $\mathbf{u}^*(t_0) = \mathbf{0}$ , while the first summand,  $\mathbf{z}(t) = e^{(t-t_0)A} \mathbf{b}$  for  $\mathbf{b} \in \mathbb{R}^n$ , constitutes the general solution to the homogeneous system.

**Example 10.36.** Our goal is to solve the initial value problem

$$\begin{aligned}\dot{u}_1 &= 2u_1 - u_2, & u_1(0) &= 1, \\ \dot{u}_2 &= 4u_1 - 3u_2 + e^t, & u_2(0) &= 0.\end{aligned}\quad (10.61)$$

The first step is to determine the eigenvalues and eigenvectors of the coefficient matrix:

$$A = \begin{pmatrix} 2 & -1 \\ 4 & -3 \end{pmatrix} \quad \text{so} \quad \lambda_1 = 1, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = -2, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

The resulting eigensolutions form the columns of the nonsingular matrix solution

$$U(t) = \begin{pmatrix} e^t & e^{-2t} \\ e^t & 4e^{-2t} \end{pmatrix}, \quad \text{hence} \quad e^{tA} = U(t)U(0)^{-1} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{4}{3}e^{-2t} \end{pmatrix}.$$

Since  $t_0 = 0$ , the two constituents of the solution formula (10.60) are

$$e^{tA} \mathbf{b} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{4}{3}e^{-2t} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} \end{pmatrix},$$

which forms the solution to the homogeneous system for the given nonzero initial conditions, and

$$\begin{aligned}\int_0^t e^{(t-s)A} \mathbf{f}(s) ds &= \int_0^t \begin{pmatrix} \frac{4}{3}e^{t-s} - \frac{1}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{1}{3}e^{-2(t-s)} \\ \frac{4}{3}e^{t-s} - \frac{4}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{4}{3}e^{-2(t-s)} \end{pmatrix} \begin{pmatrix} 0 \\ e^s \end{pmatrix} ds \\ &= \int_0^t \begin{pmatrix} -\frac{1}{3}e^t + \frac{1}{3}e^{-2t+3s} \\ -\frac{1}{3}e^t + \frac{4}{3}e^{-2t+3s} \end{pmatrix} ds = \begin{pmatrix} -\frac{1}{3}te^t + \frac{1}{9}(e^t - e^{-2t}) \\ -\frac{1}{3}te^t + \frac{4}{9}(e^t - e^{-2t}) \end{pmatrix},\end{aligned}$$

which is the particular solution to the inhomogeneous system that satisfies the homogeneous initial conditions  $\mathbf{u}(0) = \mathbf{0}$ . The solution to our initial value problem is their sum:

$$\mathbf{u}(t) = \begin{pmatrix} -\frac{1}{3}te^t + \frac{13}{9}e^t - \frac{4}{9}e^{-2t} \\ -\frac{1}{3}te^t + \frac{16}{9}e^t - \frac{16}{9}e^{-2t} \end{pmatrix}.$$

## Exercises

10.4.47. Solve the following initial value problems: (a)  $\begin{aligned}\dot{u}_1 &= 2u_1 - u_2, & u_1(0) &= 0, \\ \dot{u}_2 &= 4u_1 - 3u_2 + e^{2t}, & u_2(0) &= 0.\end{aligned}$

(b)  $\begin{aligned}\dot{u}_1 &= -u_1 + 2u_2 + e^t, & u_1(1) &= 1, \\ \dot{u}_2 &= 2u_1 - u_2 + e^t, & u_2(1) &= 1.\end{aligned}$  (c)  $\begin{aligned}\dot{u}_1 &= -u_2, & u_1(0) &= 0, \\ \dot{u}_2 &= 4u_1 + \cos t, & u_2(0) &= 1.\end{aligned}$

(d)  $\begin{aligned}\dot{u} &= 3u + v + 1, & u(1) &= 1, \\ \dot{v} &= 4u + t, & v(1) &= -1.\end{aligned}$  (e)  $\begin{aligned}\dot{p} &= p + q + t, & p(0) &= 0, \\ \dot{q} &= -p - q + t, & q(0) &= 0.\end{aligned}$

10.4.48. Solve the following initial value problems:

$$\begin{aligned}\begin{array}{lll} \dot{u}_1 = -2u_2 + 2u_3, & u_1(0) = 1, & \dot{u}_1 = u_1 - 2u_2, & u_1(0) = -1, \\ \text{(a)} \quad \dot{u}_2 = -u_1 + u_2 - 2u_3 + t, & u_2(0) = 0, & \text{(b)} \quad \dot{u}_2 = -u_2 + e^{-t}, & u_2(0) = 0, \\ \dot{u}_3 = -3u_1 + u_2 - 2u_3 + 1, & u_3(0) = 0. & \dot{u}_3 = 4u_1 - 4u_2 - u_3, & u_3(0) = -1. \end{array}\end{aligned}$$

- 10.4.49. Suppose that  $\lambda$  is *not* an eigenvalue of  $A$ . Show that the inhomogeneous system  $\dot{\mathbf{u}} = A\mathbf{u} + e^{\lambda t}\mathbf{v}$  has a solution of the form  $\mathbf{u}^*(t) = e^{\lambda t}\mathbf{w}$ , where  $\mathbf{w}$  is a constant vector. What is the general solution?

- 10.4.50. (a) Write down an integral formula for the solution to the initial value problem

$$\frac{d\mathbf{u}}{dt} = A\mathbf{u} + \mathbf{b}, \quad \mathbf{u}(0) = \mathbf{0}, \quad \text{where } \mathbf{b} \text{ is a constant vector.}$$

- (b) Suppose  $\mathbf{b} \in \text{img } A$ . Do you recover the solution you found in Exercise 10.2.24?

## 10.5 Dynamics of Structures

Chapter 6 was concerned with the equilibrium configurations of mass–spring chains and, more generally, structures constructed out of elastic bars. We are now able to undertake an analysis of their dynamical motions, which are governed by second order systems of ordinary differential equations. The same systems also serve to model the vibrations of molecules, of fundamental importance in chemistry and spectroscopy, [91]. As in the first order case, the eigenvalues of the coefficient matrix play an essential role in both the explicit solution formula and the system’s qualitative behavior(s).

Let us begin with a mass–spring chain consisting of  $n$  masses  $m_1, \dots, m_n$  connected together in a row and, possibly, to top and bottom supports by springs. As in Section 6.1, that is, we restrict our attention to purely one-dimensional motion of the masses in the direction of the chain. Thus the collective motion of the chain is prescribed by the displacement vector  $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))^T$  whose  $i^{\text{th}}$  entry represents the displacement from equilibrium of the  $i^{\text{th}}$  mass. Since we are now interested in dynamics, the displacements are allowed to depend on time,  $t$ .

The motion of each mass is subject to Newton’s Second Law:

$$\text{Force} = \text{Mass} \times \text{Acceleration}. \quad (10.62)$$

The acceleration of the  $i^{\text{th}}$  mass is the second derivative  $\ddot{u}_i = d^2u_i/dt^2$  of its displacement  $u_i(t)$ , so the right-hand sides of Newton’s Law is  $m_i \ddot{u}_i$ . These form the entries of the vector  $M \ddot{\mathbf{u}}$  obtained by multiplying the acceleration vector by the diagonal, positive definite mass matrix  $M = \text{diag}(m_1, \dots, m_n)$ . Keep in mind that the masses of the springs are assumed to be negligible in this model.

If, to begin with, we assume that there are no frictional effects, then the force exerted on each mass is the difference between the external force, if any, and the internal force due to the elongations of its two connecting springs. According to (6.11), the internal forces are the entries of the product  $K\mathbf{u}$ , where  $K = A^T C A$  is the stiffness matrix, constructed from the chain’s (reduced) incidence matrix  $A$ , and the diagonal matrix of spring constants  $C$ . Thus, Newton’s law immediately leads to the linear system of second order differential equations

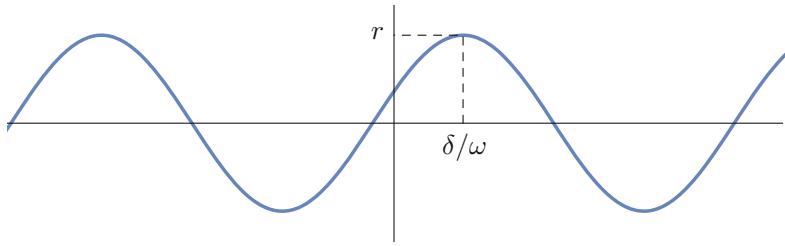
$$M \frac{d^2\mathbf{u}}{dt^2} = \mathbf{f}(t) - K\mathbf{u}, \quad (10.63)$$

governing the dynamical motions of the masses under a possibly time-dependent external force  $\mathbf{f}(t)$ . Such systems are also used to model the undamped dynamical motion of structures and molecules as well as resistanceless (superconducting) electrical circuits.

As always, the first order of business is to analyze the corresponding homogeneous system

$$M \frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} = \mathbf{0}, \quad (10.64)$$

modeling the unforced motions of the physical apparatus.



**Figure 10.5.** Vibration of a Mass.

**Example 10.37.** The simplest case is that of a single mass connected to a fixed support by a spring. Assuming no external force, the dynamical system (10.64) reduces to a homogeneous second order scalar equation

$$m \frac{d^2u}{dt^2} + k u = 0, \quad (10.65)$$

in which  $m > 0$  is the mass, while  $k > 0$  is the spring's stiffness. The general solution to (10.65) is

$$u(t) = c_1 \cos \omega t + c_2 \sin \omega t = r \cos(\omega t - \delta), \quad \text{where} \quad \omega = \sqrt{\frac{k}{m}} \quad (10.66)$$

is the natural frequency of vibration. In the second expression, we have used the phase-amplitude equation (2.7) to rewrite the solution as a single cosine with

$$\text{amplitude } r = \sqrt{c_1^2 + c_2^2} \quad \text{and phase shift } \delta = \tan^{-1} \frac{c_2}{c_1}. \quad (10.67)$$

Thus, the mass' motion is periodic, with period  $P = 2\pi/\omega$ . The stiffer the spring or the lighter the mass, the faster the vibrations. Take note of the square root in the frequency formula; quadrupling the mass slows down the vibrations by only a factor of two.

The constants  $c_1, c_2$  — or their phase-amplitude counterparts  $r, \delta$  — are determined by the initial conditions. Physically, we need to specify both an initial position and an initial velocity

$$u(t_0) = a, \quad \dot{u}(t_0) = b, \quad (10.68)$$

in order to uniquely prescribe the subsequent motion of the system. The resulting solution is most conveniently written in the form

$$u(t) = a \cos \omega(t - t_0) + \frac{b}{\omega} \sin \omega(t - t_0) = r \cos[\omega(t - t_0) - \delta],$$

with amplitude  $r = \sqrt{a^2 + \frac{b^2}{\omega^2}}$  and phase shift  $\delta = \tan^{-1} \frac{b}{a\omega}$ . (10.69)

A typical solution is plotted in [Figure 10.5](#).

Let us turn to a more general second order system. To begin with, let us assume that the masses are all the same and equal to 1 (in some appropriate units), so that (10.64) reduces to

$$\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} = \mathbf{0}. \quad (10.70)$$

Inspired by the form of the solution of the scalar equation, let us try a trigonometric ansatz for the solution, setting

$$\mathbf{u}(t) = \cos(\omega t) \mathbf{v}, \quad (10.71)$$

in which the vibrational frequency  $\omega$  is a constant scalar and  $\mathbf{v} \neq \mathbf{0}$  a constant vector. Differentiation produces

$$\frac{d\mathbf{u}}{dt} = -\omega \sin(\omega t) \mathbf{v}, \quad \frac{d^2\mathbf{u}}{dt^2} = -\omega^2 \cos(\omega t) \mathbf{v}, \quad \text{whereas} \quad K\mathbf{u} = \cos(\omega t) K\mathbf{v},$$

since the cosine factor is a scalar. Therefore, (10.71) will solve the second order system (10.70) if and only if

$$K\mathbf{v} = \omega^2 \mathbf{v}. \quad (10.72)$$

The result is in the form of the eigenvalue equation  $K\mathbf{v} = \lambda\mathbf{v}$  for the stiffness matrix  $K$ , with eigenvector  $\mathbf{v} \neq \mathbf{0}$  and eigenvalue

$$\lambda = \omega^2. \quad (10.73)$$

Now, the scalar equation has both cosine and sine solutions. By the same token, the ansatz  $\mathbf{u}(t) = \sin(\omega t) \mathbf{v}$  leads to the *same* eigenvector equation (10.72). We conclude that each positive eigenvalue leads to two different periodic trigonometric solutions.

Summarizing, we have established:

**Lemma 10.38.** If  $\mathbf{v}$  is an eigenvector of the positive definite matrix  $K$  with eigenvalue  $\lambda = \omega^2 > 0$ , then  $\mathbf{u}(t) = \cos(\omega t) \mathbf{v}$  and  $\mathbf{u}(t) = \sin(\omega t) \mathbf{v}$  are both solutions to the homogeneous second order system  $\ddot{\mathbf{u}} + K\mathbf{u} = \mathbf{0}$ .

## Stable Structures

Let us begin with the motion of a stable mass-spring chain or structure, of the type introduced in Section 6.3. According to Theorem 6.8, stability requires that the reduced stiffness matrix be positive definite:  $K > 0$ . Theorem 8.35 then says that all the eigenvalues of  $K$  are strictly positive,  $\lambda_i > 0$ , which is good, since it implies that the vibrational frequencies  $\omega_i = \sqrt{\lambda_i}$  are all real. Moreover, positive definite matrices are always complete, and so  $K$  possesses an orthogonal eigenvector basis  $\mathbf{v}_1, \dots, \mathbf{v}_n$  of  $\mathbb{R}^n$  corresponding to its eigenvalues  $\lambda_1, \dots, \lambda_n$ , listed in accordance with their multiplicities. This yields a total of  $2n$  linearly independent trigonometric eigensolutions, namely

$$\begin{aligned} \mathbf{u}_i(t) &= \cos(\omega_i t) \mathbf{v}_i = \cos(\sqrt{\lambda_i} t) \mathbf{v}_i, & i &= 1, \dots, n, \\ \tilde{\mathbf{u}}_i(t) &= \sin(\omega_i t) \mathbf{v}_i = \sin(\sqrt{\lambda_i} t) \mathbf{v}_i, \end{aligned} \quad (10.74)$$

which is precisely the number required by the general existence and uniqueness theorems for linear ordinary differential equations. The general solution to (10.70) is an arbitrary linear combination of the eigensolutions:

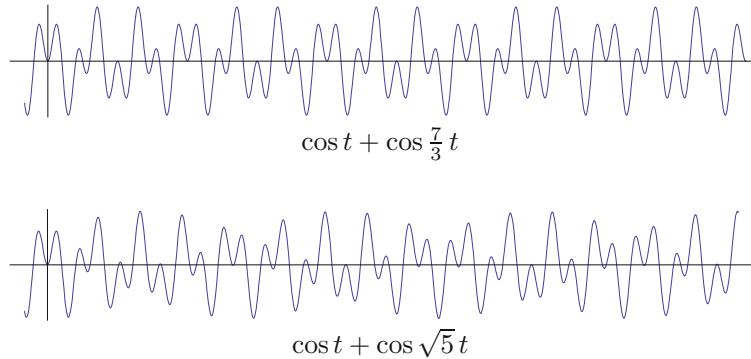
$$\mathbf{u}(t) = \sum_{i=1}^n [c_i \cos(\omega_i t) + d_i \sin(\omega_i t)] \mathbf{v}_i = \sum_{i=1}^n r_i \cos(\omega_i t - \delta_i) \mathbf{v}_i. \quad (10.75)$$

The  $2n$  coefficients  $c_i, d_i$  — or their phase-amplitude counterparts  $r_i \geq 0$  and  $0 \leq \delta_i < 2\pi$  — are uniquely determined by the initial conditions. As in (10.68), we need to specify both the initial positions and initial velocities of all the masses; this requires a total of  $2n$  initial conditions

$$\mathbf{u}(t_0) = \mathbf{a}, \quad \dot{\mathbf{u}}(t_0) = \mathbf{b}. \quad (10.76)$$

Suppose  $t_0 = 0$ ; then substituting the solution formula (10.75) into the initial conditions, we obtain

$$\mathbf{u}(0) = \sum_{i=1}^n c_i \mathbf{v}_i = \mathbf{a}, \quad \dot{\mathbf{u}}(0) = \sum_{i=1}^n \omega_i d_i \mathbf{v}_i = \mathbf{b}.$$



**Figure 10.6.** Periodic and Quasi-Periodic Functions.

Since the eigenvectors are orthogonal, the coefficients are immediately found by our orthogonal basis formula (4.7), whence

$$c_i = \frac{\langle \mathbf{a}, \mathbf{v}_i \rangle}{\|\mathbf{v}_i\|^2}, \quad d_i = \frac{\langle \mathbf{b}, \mathbf{v}_i \rangle}{\omega_i \|\mathbf{v}_i\|^2}. \quad (10.77)$$

The eigensolutions (10.74) are also known as the *normal modes of vibration* of the system, and the  $\omega_i = \sqrt{\lambda_i}$  its *natural frequencies*, which are the *square roots of the eigenvalues of the stiffness matrix  $K$* . Each eigensolution is a periodic, vector-valued function of period  $P_i = 2\pi/\omega_i$ . Linear combinations of such periodic functions are called *quasi-periodic*, because they are *not*, typically, periodic!

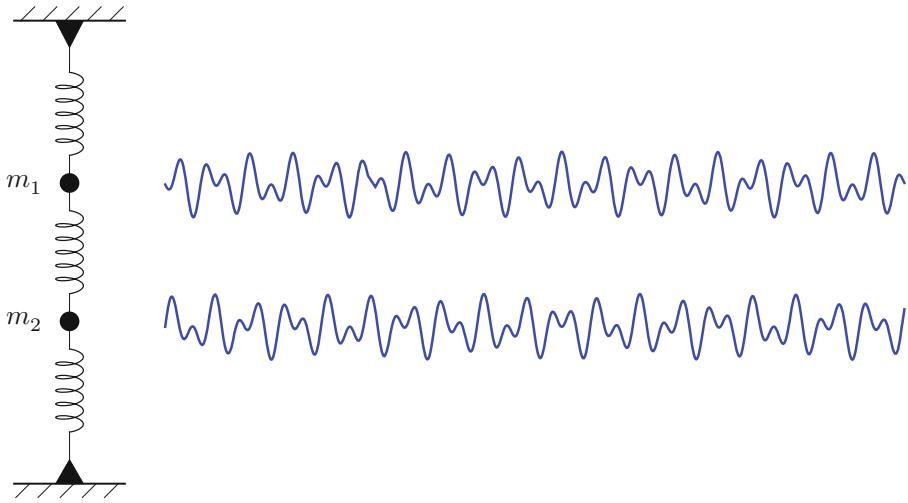
A simple example is provided by the family of functions

$$f(t) = \cos t + \cos \omega t.$$

If  $\omega = p/q \in \mathbb{Q}$  is a rational number, so  $p, q \in \mathbb{Z}$  with  $q > 0$ , then  $f(t)$  is a periodic function, since  $f(t + 2\pi q) = f(t)$ , where  $2\pi q$  is the minimal period, provided that  $p$  and  $q$  have no common factors. However, if  $\omega$  is an irrational number, then  $f(t)$  is not periodic. You are encouraged to carefully inspect the graphs in Figure 10.6. The first is periodic — can you spot where it begins to repeat? — whereas the second is only quasi-periodic and never quite succeeds in repeating its behavior. The general solution (10.75) to a vibrational system is similarly quasi-periodic, and is periodic only when *all* the frequency ratios  $\omega_i/\omega_j$  are rational numbers. To the uninitiated, such quasi-periodic motions may appear to be rather chaotic,<sup>†</sup> even though they are built from a few simple periodic constituents. Most structures and circuits exhibit quasi-periodic vibrational motions. Let us analyze a couple of simple examples.

**Example 10.39.** Consider a chain consisting of two equal unit masses connected to top and bottom supports by three springs, as in Figure 10.7, with incidence matrix  $A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$ . If the spring constants are  $c_1, c_2, c_3$  (labeled from top to bottom),

<sup>†</sup> This is *not* true chaos, which is an inherently nonlinear phenomenon, [56].



**Figure 10.7.** Motion of a Double Mass-Spring Chain with Fixed Supports.

then the stiffness matrix is

$$K = A^T C A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 + c_3 \end{pmatrix}.$$

The eigenvalues and eigenvectors of  $K$  will prescribe the normal modes and vibrational frequencies of our two-mass chain.

Let us look in detail when the springs are identical, and choose our units so that  $c_1 = c_2 = c_3 = 1$ . The resulting stiffness matrix  $K = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$  has eigenvalues and eigenvectors

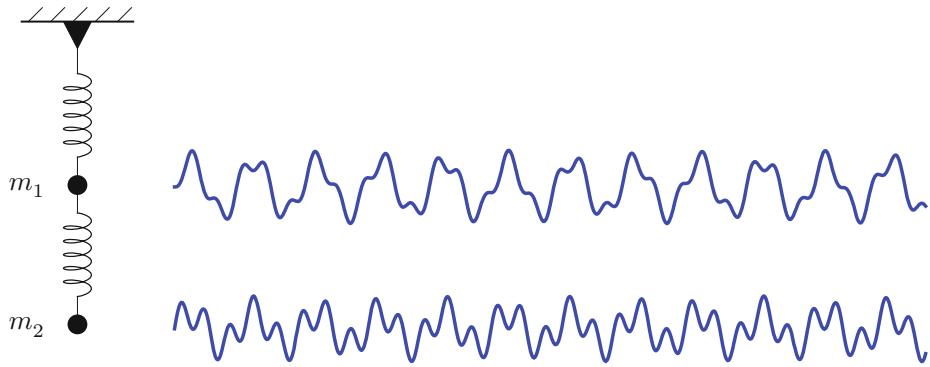
$$\lambda_1 = 1, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda_2 = 3, \quad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The general solution to the system is then

$$\mathbf{u}(t) = r_1 \cos(t - \delta_1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + r_2 \cos(\sqrt{3}t - \delta_2) \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The first summand is the normal mode vibrating at the relatively slow frequency  $\omega_1 = 1$ , with the two masses moving in tandem. The second normal mode vibrate faster, with frequency  $\omega_2 = \sqrt{3} \simeq 1.73205$ , in which the two masses move in opposing directions. The general motion is a linear combination of these two normal modes. Since the frequency ratio  $\omega_2/\omega_1 = \sqrt{3}$  is irrational, the motion is quasi-periodic. The system never quite returns to its initial configuration — unless it happens to be vibrating in one of the normal modes. A graph of some typical displacements of the masses is plotted in [Figure 10.7](#).

If we eliminate the bottom spring, so the masses are just hanging from the top support as in [Figure 10.8](#), then the reduced incidence matrix  $A^* = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$  loses its last row.



**Figure 10.8.** Motion of a Double Mass-Spring Chain with One Free End.

Assuming that the springs have unit stiffnesses  $c_1 = c_2 = 1$ , the corresponding reduced stiffness matrix is

$$K^* = (A^*)^T A^* = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors are

$$\lambda_1 = \frac{3 - \sqrt{5}}{2}, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ \frac{\sqrt{5} + 1}{2} \end{pmatrix}, \quad \lambda_2 = \frac{3 + \sqrt{5}}{2}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ -\frac{\sqrt{5} - 1}{2} \end{pmatrix}.$$

The general solution to the system is the quasi-periodic linear combination

$$\mathbf{u}(t) = r_1 \cos\left(\frac{\sqrt{5} - 1}{2}t - \delta_1\right) \begin{pmatrix} 1 \\ \frac{\sqrt{5} + 1}{2} \end{pmatrix} + r_2 \cos\left(\frac{\sqrt{5} + 1}{2}t - \delta_2\right) \begin{pmatrix} 1 \\ -\frac{\sqrt{5} - 1}{2} \end{pmatrix}.$$

The slower normal mode, with frequency  $\omega_1 = \sqrt{\frac{3 - \sqrt{5}}{2}} = \frac{\sqrt{5} - 1}{2} \simeq .61803$ , has the masses moving in tandem, with the bottom mass moving proportionally  $\frac{\sqrt{5} + 1}{2} \simeq 1.61803$  farther. The faster normal mode, with frequency  $\omega_2 = \sqrt{\frac{3 + \sqrt{5}}{2}} = \frac{\sqrt{5} + 1}{2} \simeq 1.61803$ , has the masses moving in opposing directions, with the top mass experiencing the larger displacement. Thus, removing the bottom support has caused both modes to vibrate slower. A typical solution is plotted in [Figure 10.8](#).

**Example 10.40.** Consider a three mass-spring chain, with unit springs and masses, and both ends attached to fixed supports. The stiffness matrix  $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$  has eigenvalues and eigenvectors

$$\begin{aligned} \lambda_1 &= 2 - \sqrt{2}, & \lambda_2 &= 2, & \lambda_3 &= 2 + \sqrt{2}, \\ \mathbf{v}_1 &= \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, & \mathbf{v}_2 &= \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, & \mathbf{v}_3 &= \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}. \end{aligned}$$

The three normal modes, from slowest to fastest, have frequencies

- (a)  $\omega_1 = \sqrt{2 - \sqrt{2}}$  : all three masses move in tandem, with the middle one moving  $\sqrt{2}$  times as far.
- (b)  $\omega_2 = \sqrt{2}$  : the two outer masses move in opposing directions, while the middle mass does not move.
- (c)  $\omega_3 = \sqrt{2 + \sqrt{2}}$  : the two outer masses move in tandem, while the inner mass moves  $\sqrt{2}$  times as far in the opposing direction.

The general motion is a quasi-periodic combination of these three normal modes. As such, to the naked eye it can look very complicated. Our mathematical analysis unmasks the innate simplicity, where the complex dynamics are, in fact, entirely governed by just three fundamental modes of vibration.

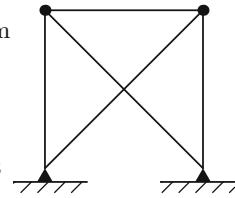
## Exercises

- 10.5.1. A 6 kilogram mass is connected to a spring with stiffness  $21 \text{ kg/sec}^2$ . Determine the frequency of vibration in hertz (cycles per second).
- 10.5.2. The lowest audible frequency is about 20 hertz = 20 cycles per second. How small a mass would need to be connected to a unit spring to produce a fast enough vibration to be audible? (As always, we assume the spring has negligible mass, which is probably not so reasonable in this situation.)
- 10.5.3. Graph the following functions. Which are periodic? quasi-periodic? If periodic, what is the (minimal) period? (a)  $\sin 4t + \cos 6t$ , (b)  $1 + \sin \pi t$ , (c)  $\cos \frac{1}{2}\pi t + \cos \frac{1}{3}\pi t$ , (d)  $\cos t + \cos \pi t$ , (e)  $\sin \frac{1}{4}t + \sin \frac{1}{5}t + \sin \frac{1}{6}t$ , (f)  $\cos t + \cos \sqrt{2}t + \cos 2t$ , (g)  $\sin t \sin 3t$ .
- 10.5.4. What is the minimal period of a function of the form  $\cos \frac{p}{q}t + \cos \frac{r}{s}t$ , assuming that each fraction is in lowest terms, i.e., its numerator and denominator have no common factors?
- 10.5.5. (a) Determine the natural frequencies of the Newtonian system  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 3 & -2 \\ -2 & 6 \end{pmatrix} \mathbf{u} = \mathbf{0}$ .  
 (b) What is the dimension of the space of solutions? Explain your answer.  
 (c) Write out the general solution. (d) For which initial conditions is the resulting motion (i) periodic? (ii) quasi-periodic? (iii) both? (iv) neither? Justify your answer.
- 10.5.6. Answer Exercise 10.5.5 for the system  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 73 & 36 \\ 36 & 52 \end{pmatrix} \mathbf{u} = \mathbf{0}$ .
- 10.5.7. Find the general solution to the following second order systems:
- (a)  $\frac{d^2u}{dt^2} = -3u + 2v, \frac{d^2v}{dt^2} = 2u - 3v$ . (b)  $\frac{d^2u}{dt^2} = -11u - 2v, \frac{d^2v}{dt^2} = -2u - 14v$ .  
 (c)  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix} \mathbf{u} = \mathbf{0}$ , (d)  $\frac{d^2\mathbf{u}}{dt^2} = \begin{pmatrix} -6 & 4 & -1 \\ 4 & -6 & 1 \\ -1 & 1 & -11 \end{pmatrix} \mathbf{u}$ .
- 10.5.8. Two masses are connected by three springs to top and bottom supports. Can you find a collection of spring constants  $c_1, c_2, c_3$  such that all vibrations are periodic?
- ♣ 10.5.9. Suppose the bottom support in the mass-spring chain in Example 10.40 is removed.  
 (a) Do you predict that the vibration rate will (i) speed up, (ii) slow down, or (iii) stay the same? (b) Verify your prediction by computing the new vibrational frequencies.  
 (c) Suppose the middle mass is displaced by a unit amount and then let go. Compute and graph the solutions in both situations. Discuss what you observe.

10.5.10. Show that a single mass that is connected to both the top and bottom supports by two springs of stiffnesses  $c_1, c_2$  will vibrate in the same manner as if it were connected to only one support by a spring with the combined stiffness  $c = c_1 + c_2$ .

- ♣ 10.5.11. (a) Describe, quantitatively and qualitatively, the normal modes of vibration for a mass–spring chain consisting of 3 unit masses, connected to top and bottom supports by unit springs. (b) Answer the same question when the bottom support is removed.
- ♡ 10.5.12. Find the vibrational frequencies for a mass–spring chain with  $n$  identical masses, connected by  $n + 1$  identical springs to both top and bottom supports. Is there any sort of limiting behavior as  $n \rightarrow \infty$ ? Hint: See Exercise 8.2.48.

- ♣ 10.5.13. Suppose the illustrated planar structure has unit masses at the nodes and the bars are all of unit stiffness. (a) Write down the system of differential equations that describes the dynamical vibrations of the structure. (b) How many independent modes of vibration are there? (c) Find numerical values for the vibrational frequencies. (d) Describe what happens when the structure vibrates in each of the normal modes. (e) Suppose the left-hand mass is displaced a unit horizontal distance. Determine the subsequent motion.



10.5.14. When does a homogeneous real first order linear system  $\dot{\mathbf{u}} = A\mathbf{u}$  have a quasi-periodic solution? What is the smallest dimension in which this can occur?

- ♣ 10.5.15. Suppose you are given  $n$  different springs. In which order should you connect them to unit masses so that the mass–spring chain vibrates the fastest? Does your answer depend upon the relative sizes of the spring constants? Does it depend upon whether the bottom mass is attached to a support or left hanging free? First try the case of three springs with spring stiffnesses  $c_1 = 1, c_2 = 2, c_3 = 3$ . Then try varying the stiffnesses. Finally, predict what will happen with 4 or 5 springs, and see whether you can make a general conjecture.

## Unstable Structures

So far, we have just dealt with the stable case, in which the stiffness matrix  $K$  is positive definite. Unstable configurations, which can admit rigid motions and/or mechanisms, will provide additional complications. The simplest is a single mass that is not attached to any spring. Since the mass experiences no restraining force, its motion is governed by the elementary second order ordinary differential equation

$$m \frac{d^2u}{dt^2} = 0. \quad (10.78)$$

The general solution is

$$u(t) = ct + d. \quad (10.79)$$

If  $c = 0$ , the mass sits at a fixed position, while when  $c \neq 0$ , it moves along a straight line with constant velocity.

More generally, suppose that the stiffness matrix  $K$  for our structure is only positive semi-definite. Each vector  $\mathbf{0} \neq \mathbf{v} \in \ker K$  represents a mode of instability of the system. Since  $K\mathbf{v} = \mathbf{0}$ , the vector  $\mathbf{v}$  is a *null eigenvector* with associated eigenvalue  $\lambda = 0$ . Lemma 10.38 provides us with two solutions to the dynamical equations (10.70) of “frequency”  $\omega = \sqrt{\lambda} = 0$ . The first,  $\mathbf{u}(t) = \cos(\omega t) \mathbf{v} \equiv \mathbf{v}$  is a constant solution, i.e., an equilibrium configuration of the system. Thus, an unstable system does not have a unique equilibrium position, since every null eigenvector  $\mathbf{v} \in \ker K$  is a constant solution. On the other hand, the second solution,  $\mathbf{u}(t) = \sin(\omega t) \mathbf{v} \equiv \mathbf{0}$ , is trivial, and so doesn’t help in constructing the requisite  $2n$  linearly independent basis solutions. To find the missing



**Figure 10.9.** A Triatomic Molecule.

solution(s), let us again argue in analogy with the scalar case (10.79), and try  $\mathbf{u}(t) = t \mathbf{v}$ . Fortunately, this works, since  $\dot{\mathbf{u}} = \mathbf{v}$ , so  $\ddot{\mathbf{u}} = \mathbf{0}$ . Also,  $K\mathbf{u} = t K\mathbf{v} = \mathbf{0}$ , and hence  $\mathbf{u}(t) = t \mathbf{v}$  solves the system  $\ddot{\mathbf{u}} + K\mathbf{u} = \mathbf{0}$ . Therefore, to each element of the kernel of the stiffness matrix — i.e., each rigid motion and mechanism — there is a two-dimensional family of solutions

$$\mathbf{u}(t) = (c t + d) \mathbf{v}. \quad (10.80)$$

When  $c = 0$ , the solution  $\mathbf{u}(t) = d \mathbf{v}$  reduces to a constant equilibrium; when  $c \neq 0$ , it is moving off to  $\infty$  with constant velocity in the null direction  $\mathbf{v}$ , and so represents an unstable mode of the system. The general solution will be a linear superposition of the vibrational modes corresponding to the positive eigenvalues and the unstable linear motions corresponding to the independent null eigenvectors.

**Remark.** If the null direction  $\mathbf{v} \in \ker K$  represents a rigid translation, then the entire structure will move in that direction. If  $\mathbf{v}$  represents an infinitesimal rotation, then, because our model is based on a linear approximation to the true nonlinear motions, the individual masses will move along straight lines, which are the tangent approximations to the circular motion that occurs in the true physical, nonlinear regime. We refer to the earlier discussion in Chapter 6 for details. Finally, if we excite a mechanism, then the masses will again follow straight lines, moving in different directions, whereas in the nonlinear real world the masses may move along much more complicated curved trajectories. For small motions, the distinction is not so important, while larger displacements, such as occur in the design of robots, platforms, and autonomous vehicles, [57, 75], will require dealing with the vastly more complicated nonlinear dynamical equations.

**Example 10.41.** Consider a system of three unit masses connected in a line by two unit springs, but not attached to any fixed supports, as illustrated in Figure 10.9. This chain could be viewed as a simplified model of an (unbent) triatomic molecule that is allowed to move only in the vertical direction. The incidence matrix is  $A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$ , and, since we are dealing with unit springs, the stiffness matrix is

$$K = A^T A = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors of  $K$  are easily found:

$$\lambda_1 = 0, \quad \lambda_2 = 1, \quad \lambda_3 = 3, \\ \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Each positive eigenvalue provides two trigonometric solutions, while the zero eigenvalue leads to solutions that are constant or depend linearly on  $t$ . This yields the required six basis solutions:

$$\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{u}_3(t) = \begin{pmatrix} \cos t \\ 0 \\ -\cos t \end{pmatrix}, \quad \mathbf{u}_5(t) = \begin{pmatrix} \cos \sqrt{3}t \\ -2 \cos \sqrt{3}t \\ \cos \sqrt{3}t \end{pmatrix}, \\ \mathbf{u}_2(t) = \begin{pmatrix} t \\ t \\ t \end{pmatrix}, \quad \mathbf{u}_4(t) = \begin{pmatrix} \sin t \\ 0 \\ -\sin t \end{pmatrix}, \quad \mathbf{u}_6(t) = \begin{pmatrix} \sin \sqrt{3}t \\ -2 \sin \sqrt{3}t \\ \sin \sqrt{3}t \end{pmatrix}.$$

The first solution  $\mathbf{u}_1(t)$  is a constant, equilibrium mode, where the masses rest at a fixed common distance from their reference positions. The second solution  $\mathbf{u}_2(t)$  is the unstable mode, corresponding to a uniform rigid translation of the molecule that does not stretch the interconnecting springs. The final four solutions represent vibrational modes. In the first pair,  $\mathbf{u}_3(t), \mathbf{u}_4(t)$ , the two outer masses move in opposing directions, while the middle mass remains fixed, while the final pair,  $\mathbf{u}_5(t), \mathbf{u}_6(t)$  has the two outer masses moving in tandem, while the inner mass moves twice as far in the opposite direction. The general solution is a linear combination of the six normal modes,

$$\mathbf{u}(t) = c_1 \mathbf{u}_1(t) + \cdots + c_6 \mathbf{u}_6(t), \quad (10.81)$$

and corresponds to the entire molecule moving at a fixed velocity while the individual masses perform a quasi-periodic vibration.

Let us see whether we can predict the motion of the molecule from its initial conditions

$$\mathbf{u}(0) = \mathbf{a}, \quad \dot{\mathbf{u}}(0) = \mathbf{b},$$

where  $\mathbf{a} = (a_1, a_2, a_3)^T$  indicates the initial displacements of the three atoms, while  $\mathbf{b} = (b_1, b_2, b_3)^T$  are their initial velocities. Substituting the solution formula (10.81) leads to the two linear systems

$$c_1 \mathbf{v}_1 + c_3 \mathbf{v}_2 + c_5 \mathbf{v}_3 = \mathbf{a}, \quad c_2 \mathbf{v}_1 + c_4 \mathbf{v}_2 + \sqrt{3} c_6 \mathbf{v}_3 = \mathbf{b},$$

for the coefficients  $c_1, \dots, c_6$ . As in (10.77), we can use the orthogonality of the eigenvectors to immediately compute the coefficients:

$$c_1 = \frac{\mathbf{a} \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} = \frac{a_1 + a_2 + a_3}{3}, \quad c_3 = \frac{\mathbf{a} \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} = \frac{a_1 - a_3}{2}, \quad c_5 = \frac{\mathbf{a} \cdot \mathbf{v}_3}{\|\mathbf{v}_3\|^2} = \frac{a_1 - 2a_2 + a_3}{6}, \\ c_2 = \frac{\mathbf{b} \cdot \mathbf{v}_1}{\|\mathbf{v}_1\|^2} = \frac{b_1 + b_2 + b_3}{3}, \quad c_4 = \frac{\mathbf{b} \cdot \mathbf{v}_2}{\|\mathbf{v}_2\|^2} = \frac{b_1 - b_3}{2}, \quad c_6 = \frac{\mathbf{b} \cdot \mathbf{v}_3}{\sqrt{3} \|\mathbf{v}_3\|^2} = \frac{b_1 - 2b_2 + b_3}{6\sqrt{3}}.$$

In particular, the unstable translational mode is excited if and only if  $c_2 \neq 0$ , and this occurs if and only if there is a nonzero net initial velocity of the molecule:  $b_1 + b_2 + b_3 \neq 0$ . In this case, the vibrating molecule will run off to  $\infty$  at a uniform velocity  $c = c_2 = \frac{1}{3}(b_1 + b_2 + b_3)$  equal to the average of the individual initial velocities. On the other hand, if  $b_1 + b_2 + b_3 = 0$ ,

then the atoms will vibrate quasi-periodically, with frequencies 1 and  $\sqrt{3}$ , around its fixed center of mass.

The observations established in this example hold, in fact, in complete generality. Let us state the result, leaving the details of the proof as an exercise for the reader.

**Theorem 10.42.** The general solution to an unstable second order linear system  $\ddot{\mathbf{u}} + K\mathbf{u} = \mathbf{0}$  with positive semi-definite coefficient matrix  $K \geq 0$  is a linear combination of a quasi-periodic or periodic vibrations and a uniform linear motion at a fixed velocity in the direction of a null eigenvector  $\mathbf{v} \in \ker K$ . In particular, the system will just vibrate around a fixed position if and only if the initial velocity  $\dot{\mathbf{u}}(t_0) \in (\ker K)^\perp = \text{img } K$  lies in the image of the coefficient matrix.

As in Chapter 6, the unstable modes  $\mathbf{v} \in \ker K$  correspond to either rigid motions or to mechanisms of the structure. Thus, to prevent a structure from exhibiting an unstable motion, one has to ensure that the initial velocity is orthogonal to all of the unstable modes. (The value of the initial position is not an issue.) This is the dynamical counterpart of the requirement that an external force be orthogonal to all unstable modes in order to maintain equilibrium in the structure, as in Theorem 6.8.

### Systems with Differing Masses

When a chain or structure has different masses at the nodes, the (unforced) Newtonian equations of motion take the more general form

$$M \frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} = \mathbf{0}, \quad \text{or, equivalently,} \quad \frac{d^2\mathbf{u}}{dt^2} = -M^{-1}K\mathbf{u} = -P\mathbf{u}. \quad (10.82)$$

The mass matrix  $M$  is always positive definite (and, almost always, diagonal, although this is not required by the general theory), while the stiffness matrix  $K = A^T C A$  is either positive definite or, in the unstable situation when  $\ker A \neq \{\mathbf{0}\}$ , positive semi-definite. The coefficient matrix

$$P = M^{-1}K = M^{-1}A^T C A \quad (10.83)$$

is *not* in general symmetric, and so we cannot directly apply the preceding constructions. However,  $P$  does have the more general self-adjoint form (7.85) based on the weighted inner products

$$\langle \mathbf{u}, \tilde{\mathbf{u}} \rangle = \mathbf{u}^T M \tilde{\mathbf{u}}, \quad \langle \langle \mathbf{v}, \tilde{\mathbf{v}} \rangle \rangle = \mathbf{v}^T C \tilde{\mathbf{v}}, \quad (10.84)$$

on, respectively, the domain and codomain of the (reduced) incidence matrix  $A$ . Moreover, in the stable case when  $\ker A = \{\mathbf{0}\}$ , the matrix  $P$  is positive definite in the generalized sense of Definition 7.59.

To solve the system of differential equations, we substitute the same trigonometric solution ansatz  $\mathbf{u}(t) = \cos(\omega t) \mathbf{v}$ . This results in a *generalized eigenvalue equation*

$$K\mathbf{v} = \lambda M\mathbf{v}, \quad \text{or, equivalently,} \quad P\mathbf{v} = \lambda\mathbf{v}, \quad \text{with} \quad \lambda = \omega^2. \quad (10.85)$$

The matrix  $M$  assumes the role of the identity matrix in the standard eigenvalue equation (8.13), and  $\lambda$  is a generalized eigenvalue if and only if it satisfies the generalized characteristic equation

$$\det(K - \lambda M) = 0. \quad (10.86)$$

According to Exercise 8.5.8, if  $M > 0$  and  $K > 0$ , then all the generalized eigenvalues are real and non-negative. Moreover the generalized eigenvectors form an orthogonal basis of

$\mathbb{R}^n$ , but now with respect to the weighted inner product (10.84) prescribed by the mass matrix  $M$ . The general solution is a quasi-periodic linear combination of the eigensolutions, of the same form as in (10.75). In the unstable case, when  $K \geq 0$  (but  $M$  necessarily remains positive definite), one must include enough generalized null eigenvectors to span  $\text{ker } K$ , each of which leads to an unstable mode of the form (10.80). Further details are relegated to the exercises.

## Exercises

10.5.16. Find the general solution to the following systems. Distinguish between the vibrational and unstable modes. What constraints on the initial conditions ensure

that the unstable modes are not excited? (a)  $\frac{d^2u}{dt^2} = -4u - 2v$ ,  $\frac{d^2v}{dt^2} = -2u - v$ .

(b)  $\frac{d^2u}{dt^2} = -u - 3v$ ,  $\frac{d^2v}{dt^2} = -3u - 9v$ . (c)  $\frac{d^2u}{dt^2} = -2u + v - 2w$ ,  $\frac{d^2v}{dt^2} = u - v$ ,

$\frac{d^2w}{dt^2} = -2u - 4w$ . (d)  $\frac{d^2u}{dt^2} = -u + v - 2w$ ,  $\frac{d^2v}{dt^2} = u - v + 2w$ ,  $\frac{d^2w}{dt^2} = -2u + 2v - 4w$ .

10.5.17. Let  $K = \begin{pmatrix} 3 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 3 \end{pmatrix}$ . (a) Find an orthogonal matrix  $Q$  and a diagonal matrix  $\Lambda$

such that  $K = Q \Lambda Q^T$ . (b) Is  $K$  positive definite? (c) Solve the second order system

$\frac{d^2\mathbf{u}}{dt^2} = A\mathbf{u}$  subject to the initial conditions  $\mathbf{u}(0) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ ,  $\frac{d\mathbf{u}}{dt}(0) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ .

(d) Is your solution periodic? If your answer is yes, indicate the period.

(e) Is the general solution to the system periodic?

10.5.18. Answer Exercise 10.5.17 when  $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ .

10.5.19. Compare the solutions to the mass-spring system (10.65) with tiny spring constant  $k = \varepsilon \ll 1$  to those of the completely unrestrained system (10.78). Are they close? Discuss.

◇ 10.5.20. Discuss the three-dimensional motions of the triatomic molecule of Example 10.41. Are the vibrational frequencies the same as those of the one-dimensional model?

◇ 10.5.21. So far, our mass-spring chains have been allowed to move only in the vertical direction. (a) Set up the system governing the planar motions of a mass-spring chain consisting of two unit masses attached to top and bottom supports by unit springs, where the masses are allowed to move in the longitudinal and transverse directions. Compare the resulting vibrational frequencies with the one-dimensional case. (b) Repeat the analysis when the bottom support is removed. (c) Can you make any conjectures concerning the planar motions of general mass-spring chains?

♣ 10.5.22. Find the vibrational frequencies and instabilities of the following structures, assuming they have unit masses at all the nodes. Explain in detail how each normal mode moves the structure: (a) the three bar planar structure in Figure 6.13; (b) its reinforced version in Figure 6.16; (c) the swing set in Figure 6.18.

♣ 10.5.23. Assuming unit masses at the nodes, find the vibrational frequencies and describe the normal modes for the following planar structures. What initial conditions will not excite its instabilities? (a) An equilateral triangle; (b) a square; (c) a regular hexagon.

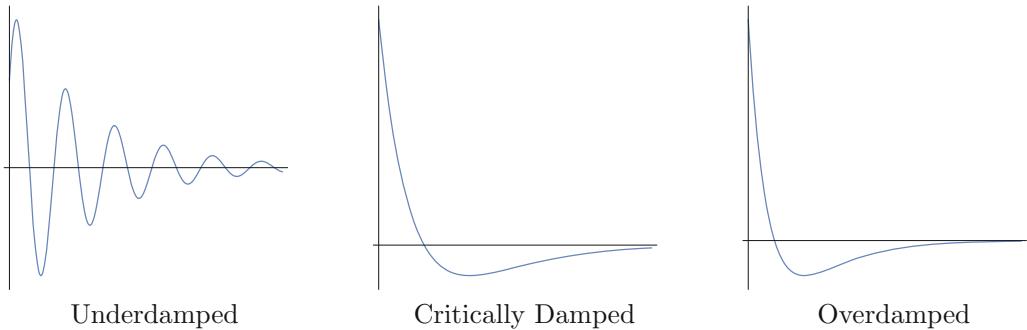
♣ 10.5.24. Answer Exercise 10.5.23 for the three-dimensional motions of a regular tetrahedron.

- ◇ 10.5.25. (a) Show that if a structure contains all unit masses and bars with unit stiffness,  $c_i = 1$ , then its frequencies of vibration are the nonzero singular values of the reduced incidence matrix. (b) How would you recognize when a structure is close to being unstable?
- 10.5.26. Prove that if the initial velocity satisfies  $\dot{\mathbf{u}}(t_0) = \mathbf{b} \in \text{coimg } A$ , then the solution to the initial value problem (10.70, 76) remains bounded.
- 10.5.27. Find the general solution to the system (10.82) for the following matrix pairs:
- (a)  $M = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ ,  $K = \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$ , (b)  $M = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix}$ ,  $K = \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix}$ ,
- (c)  $M = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ ,  $K = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ , (d)  $M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{pmatrix}$ ,  $K = \begin{pmatrix} 5 & -1 & -1 \\ -1 & 6 & 3 \\ -1 & 3 & 9 \end{pmatrix}$ ,
- (e)  $M = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ ,  $K = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$ , (f)  $M = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$ ,  $K = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 8 & 2 \\ 0 & 2 & 1 \end{pmatrix}$ .
- 10.5.28. A mass-spring chain consists of two masses,  $m_1 = 1$  and  $m_2 = 2$ , connected to top and bottom supports by identical springs with unit stiffness. The upper mass is displaced by a unit distance. Find the subsequent motion of the system.
- 10.5.29. Answer Exercise 10.5.28 when the bottom support is removed.
- ♣ 10.5.30. (a) A water molecule consists of two hydrogen atoms connected at an angle of  $105^\circ$  to an oxygen atom whose relative mass is 16 times that of each of the hydrogen atoms. If the molecular bonds are modeled as linear unit springs, determine the fundamental frequencies and describe the corresponding vibrational modes. (b) Do the same for a carbon tetrachloride molecule, in which the chlorine atoms, with atomic weight 35, are positioned on the vertices of a regular tetrahedron and the carbon atom, with atomic weight 12, is at the center. (c) Finally try a benzene molecule, consisting of 6 carbon atoms arranged in a regular hexagon. In this case, every other bond is double strength because two electrons are shared. (Ignore the six extra hydrogen atoms for simplicity.)
- ◇ 10.5.31. Repeat Exercise 10.5.21 for fully 3-dimensional motions of the chain.
- ♠ 10.5.32. Suppose you have masses  $m_1 = 1$ ,  $m_2 = 2$ ,  $m_3 = 3$ , connected to top and bottom supports by identical unit springs. Does rearranging the order of the masses change the fundamental frequencies? If so, which order produces the fastest vibrations?
- ◇ 10.5.33. Suppose  $M$  is a nonsingular matrix. Prove that  $\lambda$  is a generalized eigenvalue of the matrix pair  $K, M$  if and only if it is an ordinary eigenvalue of the matrix  $P = M^{-1}K$ . How are the eigenvectors related? How are the characteristic equations related?
- 10.5.34. Suppose that  $\mathbf{u}(t)$  is a solution to (10.82). Let  $N = \sqrt{M}$  denote the positive definite square root of the mass matrix  $M$ , as defined in Exercise 8.5.27. (a) Prove that the “weighted” displacement vector  $\tilde{\mathbf{u}}(t) = N\mathbf{u}(t)$  solves  $d^2\tilde{\mathbf{u}}/dt^2 = -\tilde{K}\tilde{\mathbf{u}}$ , where  $\tilde{K} = N^{-1}KN^{-1}$  is a symmetric, positive semi-definite matrix. (b) Explain in what sense this can serve as an alternative to the generalized eigenvector solution method.
- ◇ 10.5.35. Provide the details of the proof of Theorem 10.42.
- ◇ 10.5.36. State and prove the counterpart of Theorem 10.42 for the variable mass system (10.82).

---

## Friction and Damping

We have not yet allowed friction to affect the motion of our dynamical equations. In the standard physical model, the frictional force on a mass in motion is directly proportional



**Figure 10.10.** Damped Vibrations.

to its velocity, [31]. For the simplest case of a single mass attached to a spring, one amends the balance of forces in the undamped Newtonian equation (10.65) to obtain

$$m \frac{d^2u}{dt^2} + \beta \frac{du}{dt} + k u = 0. \quad (10.87)$$

As before,  $m > 0$  is the mass, and  $k > 0$  the spring stiffness, while  $\beta > 0$  measures the effect of a velocity-dependent frictional force — the larger the value of  $\beta$ , the greater the frictional force.

The solution of this more general second order homogeneous linear ordinary differential equation is found by substituting the usual exponential ansatz  $u(t) = e^{\lambda t}$ , reducing it to the quadratic characteristic equation

$$m \lambda^2 + \beta \lambda + k = 0. \quad (10.88)$$

Assuming that  $m, \beta, k > 0$ , there are three possible cases:

*Underdamped:* If  $0 < \beta < 2\sqrt{mk}$ , then (10.88) has two complex-conjugate roots:

$$\lambda = -\frac{\beta}{2m} \pm i \frac{\sqrt{4mk - \beta^2}}{2m} = -\mu \pm i\nu. \quad (10.89)$$

The general solution to the differential equation,

$$u(t) = e^{-\mu t} (c_1 \cos \nu t + c_2 \sin \nu t) = r e^{-\mu t} \cos(\nu t - \delta), \quad (10.90)$$

represents a damped periodic motion. The mass continues to oscillate at a fixed frequency

$$\nu = \frac{\sqrt{4mk - \beta^2}}{2m} = \sqrt{\frac{k}{m} - \frac{\beta^2}{4m^2}}, \quad (10.91)$$

but the vibrational amplitude  $r e^{-\mu t}$  decays to zero at an exponential rate as  $t \rightarrow \infty$ . Observe that, in a rigorous mathematical sense, the mass never quite returns to equilibrium, although in the real world, after a sufficiently long time the residual vibrations are not noticeable, and equilibrium is physically (but not mathematically) achieved. The rate of decay,  $\mu = \beta/(2m)$ , is directly proportional to the friction, and inversely proportional to the mass. Thus, greater friction and/or less mass will accelerate the return to equilibrium. The friction also has an effect on the vibrational frequency (10.91); the larger  $\beta$  is, the slower the oscillations become and the more rapid the damping effect. As the friction approaches the critical threshold  $\beta_* = 2\sqrt{mk}$ , the vibrational frequency goes to zero,  $\nu \rightarrow 0$ , and so the oscillatory period  $2\pi/\nu$  becomes longer and longer.

*Overdamped:* If  $\beta > 2\sqrt{mk}$ , then the characteristic equation (10.88) has two negative real roots

$$\lambda_1 = -\frac{\beta + \sqrt{\beta^2 - 4mk}}{2m} < \lambda_2 = -\frac{\beta - \sqrt{\beta^2 - 4mk}}{2m} < 0.$$

The solution

$$u(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t} \quad (10.92)$$

is a linear combination of two decaying exponentials. An overdamped system models the motion of, say, a mass in a vat of molasses. Its “vibration” is so slow that it can pass at most once through the equilibrium position, and then only when its initial velocity is relatively large. In the long term, the first exponential in the solution (10.92) will go to zero faster, and hence, as long as  $c_2 \neq 0$ , the overall decay rate of the solution is governed by the dominant (least negative) eigenvalue  $\lambda_2$ .

*Critically Damped:* The borderline case occurs when  $\beta = \beta_* = 2\sqrt{mk}$ , which means that the characteristic equation (10.88) has only a single negative real root:

$$\lambda_1 = -\frac{\beta}{2m}.$$

In this case, our ansatz supplies only one exponential solution  $e^{\lambda_1 t} = e^{-\beta t/(2m)}$ . A second independent solution is obtained by multiplication by  $t$ , leading to the general solution

$$u(t) = (c_1 t + c_2) e^{-\beta t/(2m)}. \quad (10.93)$$

Even though the formula looks quite different, its qualitative behavior is very similar to the overdamped case. The factor  $t$  plays an unimportant role, since the asymptotics of this solution are almost entirely governed by the decaying exponential function. This represents a non-vibrating solution that has the slowest possible decay rate, since any further reduction of the frictional coefficient will allow a damped, slowly oscillatory vibration to appear.

In all three cases, the zero equilibrium solution is globally asymptotically stable. Physically, no matter how small the frictional contribution, all solutions to the unforced system eventually return to equilibrium as friction eventually overwhelms the motion.

This concludes our discussion of the scalar case. Similar considerations apply to mass–spring chains, and to two- and three-dimensional structures. A frictionally damped structure is modeled by a second order system of the form

$$M \frac{d^2 \mathbf{u}}{dt^2} + B \frac{d\mathbf{u}}{dt} + K \mathbf{u} = \mathbf{0}, \quad (10.94)$$

where the mass matrix  $M$  and the matrix of frictional coefficients  $B$  are both diagonal and positive definite, while the stiffness matrix  $K = A^T C A \geq 0$  is a positive semi-definite Gram matrix constructed from the (reduced) incidence matrix  $A$ . Under these assumptions, it can be proved that the zero equilibrium solution is globally asymptotically stable. However, the mathematical details in this case are sufficiently intricate that we shall leave their analysis as an advanced project for the highly motivated student.

## Exercises

- 10.5.37. Consider the overdamped mass–spring equation  $\ddot{u} + 6\dot{u} + 5u = 0$ . If the mass starts out a distance 1 away from equilibrium, how large must the initial velocity be in order that it pass through equilibrium once?

10.5.38. Solve the following mass-spring initial value problems, and classify as to

(i) overdamped, (ii) critically damped, (iii) underdamped, or (iv) undamped:

- (a)  $\ddot{u} + 6\dot{u} + 9u = 0$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 1$ . (b)  $\ddot{u} + 2\dot{u} + 10u = 0$ ,  $u(0) = 1$ ,  $\dot{u}(0) = 1$ .  
 (c)  $\ddot{u} + 16u = 0$ ,  $u(1) = 0$ ,  $\dot{u}(1) = 1$ . (d)  $\ddot{u} + 3\dot{u} + 9u = 0$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 1$ .  
 (e)  $2\ddot{u} + 3\dot{u} + u = 0$ ,  $u(0) = 2$ ,  $\dot{u}(0) = 0$ . (f)  $\ddot{u} + 6\dot{u} + 10u = 0$ ,  $u(0) = 3$ ,  $\dot{u}(0) = -2$ .

10.5.39. (a) A mass weighing 16 pounds stretches a spring 6.4 feet. Assuming no friction, determine the equation of motion and the natural frequency of vibration of the mass-spring system. Use the value  $g = 32 \text{ ft/sec}^2$  for the gravitational acceleration. (b) The mass-spring system is placed in a jar of oil, whose frictional resistance equals the speed of the mass. Assume the spring is stretched an additional 2 feet from its equilibrium position and let go. Determine the motion of the mass. (c) Is the system overdamped or underdamped? Are the vibrations more rapid or less rapid than in the undamped system?

10.5.40. Suppose you convert the second order equation (10.87) into its phase plane equivalent. What are the phase portraits corresponding to (a) undamped, (b) underdamped, (c) critically damped, and (d) overdamped motion?

◇ 10.5.41. (a) Prove that, given a non-constant solution to an overdamped mass-spring system, there is at most one time where  $u(t_*) = 0$ . (b) Is this statement also valid in the critically damped case?

10.5.42. Discuss the possible behaviors of a mass moving in a frictional medium that is not attached to a spring, i.e., set  $k = 0$  in (10.87).

## 10.6 Forcing and Resonance

Up until now, our physical system has been left free to vibrate on its own. Let us investigate what happens when we shake it. In this section, we will consider the effects of periodic external forcing on both undamped and damped systems.

The simplest case is that of a single mass connected to a spring that has no frictional damping. We append an external forcing function  $f(t)$  to the homogeneous (unforced) equation (10.65), leading to the inhomogeneous second order equation

$$m \frac{d^2u}{dt^2} + k u = f(t), \quad (10.95)$$

in which  $m > 0$  is the mass and  $k > 0$  the spring stiffness. We are particularly interested in the case of periodic forcing

$$f(t) = \alpha \cos \gamma t \quad (10.96)$$

of frequency  $\gamma > 0$  and amplitude  $\alpha$ . To find a particular solution to (10.95–96), we use the method of undetermined coefficients<sup>†</sup> which tells us to guess a trigonometric solution ansatz of the form

$$u^*(t) = a \cos \gamma t + b \sin \gamma t, \quad (10.97)$$

where  $a, b$  are constants to be determined. Substituting (10.97) into the differential equation produces

$$m \frac{d^2u^*}{dt^2} + k u^* = a(k - m\gamma^2) \cos \gamma t + b(k - m\gamma^2) \sin \gamma t = \alpha \cos \gamma t.$$

<sup>†</sup> One can also use variation of parameters, although the intervening calculations are slightly more complicated.

We can solve for

$$a = \frac{\alpha}{k - m\gamma^2} = \frac{\alpha}{m(\omega^2 - \gamma^2)}, \quad b = 0, \quad (10.98)$$

where

$$\omega = \sqrt{\frac{k}{m}} \quad (10.99)$$

refers to the natural, unforced vibrational frequency of the system. The solution (10.98) is valid provided its denominator is nonzero:

$$k - m\gamma^2 = m(\omega^2 - \gamma^2) \neq 0.$$

Therefore, as long as the forcing frequency is *not* equal to the system's natural frequency, i.e.,  $\gamma \neq \omega$ , there exists a particular solution

$$u^*(t) = a \cos \gamma t = \frac{\alpha}{m(\omega^2 - \gamma^2)} \cos \gamma t \quad (10.100)$$

that vibrates at the same frequency as the forcing function.

The general solution to the inhomogeneous system (10.95) is found, as usual, by adding in an arbitrary solution (10.66) to the homogeneous equation, yielding

$$u(t) = \frac{\alpha}{m(\omega^2 - \gamma^2)} \cos \gamma t + r \cos(\omega t - \delta), \quad (10.101)$$

where  $r$  and  $\delta$  are determined by the initial conditions. The solution is therefore a quasi-periodic combination of two simple periodic motions — the second, vibrating with frequency  $\omega$ , represents the internal or natural vibrations of the system, while the first, with frequency  $\gamma$ , represents the response to the periodic forcing. Due to the factor  $\omega^2 - \gamma^2$  in the denominator of the latter, the closer the forcing frequency is to the natural frequency, the larger the overall amplitude of the response.

Suppose we start the mass at equilibrium at the initial time  $t_0 = 0$ , so the initial conditions are

$$u(0) = 0, \quad \dot{u}(0) = 0. \quad (10.102)$$

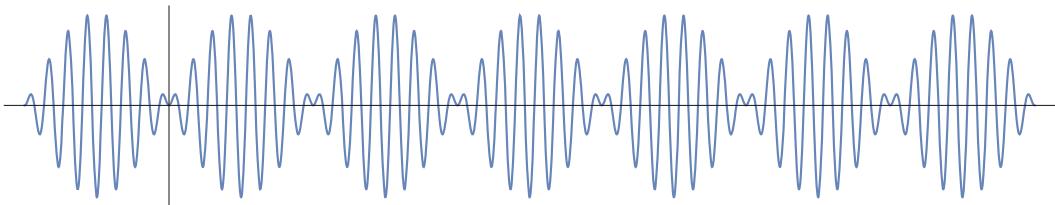
Substituting (10.101) and solving for  $r$ ,  $\delta$ , we find that

$$r = -\frac{\alpha}{m(\omega^2 - \gamma^2)}, \quad \delta = 0.$$

Thus, the solution to the initial value problem can be written in the form

$$u(t) = \frac{\alpha}{m(\omega^2 - \gamma^2)} (\cos \gamma t - \cos \omega t) = \frac{2\alpha}{m(\omega^2 - \gamma^2)} \sin\left(\frac{\omega + \gamma}{2}t\right) \sin\left(\frac{\omega - \gamma}{2}t\right), \quad (10.103)$$

where we have employed a standard trigonometric identity, cf. Exercise 3.6.17. The first trigonometric factor,  $\sin \frac{1}{2}(\omega + \gamma)t$ , represents a periodic motion at a frequency equal to the average of the natural and the forcing frequencies. If the forcing frequency  $\gamma$  is close to the natural frequency  $\omega$ , then the second factor,  $\sin \frac{1}{2}(\omega - \gamma)t$ , has a much smaller frequency, and so oscillates on a much longer time scale. As a result, it *modulates* the amplitude of the more rapid vibrations, and is responsible for the phenomenon of *beats*, in which a rapid vibration is subject to a slowly varying amplitude. An everyday illustration of beats is two tuning forks that have nearby pitches. When they vibrate close to each other, the sound



**Figure 10.11.** Beats in a Periodically Forced Vibration.

you hear waxes and wanes in intensity. As a mathematical example, [Figure 10.11](#) displays the graph of the particular function

$$\cos 14t - \cos 16t = 2 \sin t \sin 15t$$

on the interval  $-\pi \leq t \leq 6\pi$ . The slowly varying amplitude  $2 \sin t$  is clearly visible as the envelope of the relatively rapid vibrations at frequency 15.

When we force the system at exactly the natural frequency  $\gamma = \omega$ , the trigonometric ansatz (10.97) no longer works. This is because both terms are now solutions to the homogeneous equation, and so cannot be combined to form a solution to the inhomogeneous version. In this situation, there is a simple modification to the ansatz, namely multiplication by  $t$ , that does the trick. Substituting

$$u^*(t) = at \cos \omega t + bt \sin \omega t \quad (10.104)$$

into the differential equation (10.95), we obtain

$$m \frac{d^2 u^*}{dt^2} + k u^* = -2am\omega \sin \omega t + 2bm\omega \cos \omega t = \alpha \cos \omega t,$$

provided

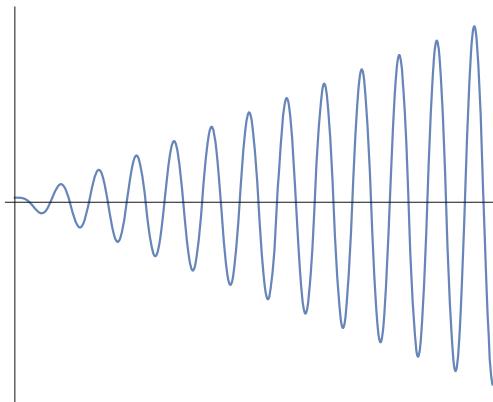
$$a = 0, \quad b = \frac{\alpha}{2m\omega}, \quad \text{and so} \quad u^*(t) = \frac{\alpha}{2m\omega} t \sin \omega t.$$

Combining the resulting particular solution with the solution to the homogeneous equation leads to the general solution

$$u(t) = \frac{\alpha}{2m\omega} t \sin \omega t + r \cos(\omega t - \delta). \quad (10.105)$$

Both terms vibrate with frequency  $\omega$ , but the amplitude of the first grows larger and larger as  $t \rightarrow \infty$ . As illustrated in [Figure 10.12](#), the mass will oscillate more and more wildly. In this situation, the system is said to be in *resonance*, and the increasingly large oscillations are provoked by forcing it at its natural frequency  $\omega$ . In a physical apparatus, once the amplitude of resonant vibrations stretches the spring beyond its elastic limits, the linear Hooke's Law model is no longer applicable, and either the spring breaks or the system enters a nonlinear regime.

Furthermore, if we are very close to resonance, the oscillations induced by the particular solution (10.103) will have extremely large, although bounded, amplitude. The lesson is, never force a system at or close to its natural frequency (or frequencies) of vibration. A classic example was the 1831 collapse of a bridge when a British infantry regiment marched in unison across it, apparently inducing a resonant vibration of the structure. The bridge in question was an early example of the suspension style, similar to that pictured in [Figure 10.13](#). Learning their lesson, soldiers nowadays no longer march in step across



**Figure 10.12.** Resonance.



**Figure 10.13.** The Albert Bridge in London.

bridges — as reminded by the sign in the photo in Figure 10.13. An even more dramatic case is the 1940 Tacoma Narrows Bridge disaster, when the vibrations due to a strong wind caused the bridge to oscillate wildly and break apart! The collapse was caught on film, which can be found on YouTube, and is extremely impressive. The traditional explanation was the excitement of the bridge's resonant frequencies, although later studies revealed a more sophisticated mathematical explanation of the collapse, [22; p. 118]. But resonance is not exclusively harmful. In a microwave oven, the electromagnetic waves are tuned to the resonant frequencies of water molecules so as to excite them into large vibrations and thereby heat up your dinner. Blowing into a clarinet or other wind instrument excites the resonant frequencies in the column of air contained within it, and this produces the musical sound vibrations that we hear.

Frictional effects can partially mollify the extreme behavior near the resonant frequency. The frictionally damped vibrations of a mass on a spring, when subject to periodic forcing,

are described by the inhomogeneous differential equation

$$m \frac{d^2u}{dt^2} + \beta \frac{du}{dt} + k u = \alpha \cos \gamma t. \quad (10.106)$$

Let us assume that the friction is sufficiently small so as to be in the underdamped regime  $\beta < 2\sqrt{mk}$ . Since neither summand solves the homogeneous system, we can use the trigonometric solution ansatz (10.97) to construct the particular solution

$$u^*(t) = \frac{\alpha}{\sqrt{m^2(\omega^2 - \gamma^2)^2 + \beta^2\gamma^2}} \cos(\gamma t - \varepsilon), \quad \text{where} \quad \omega = \sqrt{\frac{k}{m}} \quad (10.107)$$

continues to denote the undamped resonant frequency (10.99), while  $\varepsilon$ , defined by

$$\tan \varepsilon = \frac{\beta \gamma}{m(\omega^2 - \gamma^2)}, \quad (10.108)$$

represents a frictionally induced *phase lag*. Thus, the larger the friction  $\beta$ , the more pronounced the phase lag  $\varepsilon$  in the response of the system to the external forcing. As the forcing frequency  $\gamma$  increases, so does the phase lag, which attains the value  $\frac{1}{2}\pi$  at the resonant frequency  $\gamma = \omega$ , meaning that the system lags a quarter period behind the forcing, and converges to its maximum  $\varepsilon = \pi$  as  $\gamma \rightarrow \infty$ . Thus, the response to a very high frequency forcing is almost exactly out of phase — the mass is moving downwards when the force is pulling it upwards, and vice versa! The amplitude of the persistent response (10.107) is at a maximum at the resonant frequency  $\gamma = \omega$ , where it takes the value  $\alpha/(\beta\omega)$ . Thus, the smaller the frictional coefficient  $\beta$  (or the slower the resonant frequency  $\omega$ ), the more likely the breakdown of the system due to an overly large response.

The general solution is

$$u(t) = \frac{\alpha}{\sqrt{m^2(\omega^2 - \gamma^2)^2 + \beta^2\gamma^2}} \cos(\gamma t - \varepsilon) + r e^{-\mu t} \cos(\nu t - \delta), \quad (10.109)$$

where  $\lambda = \mu \pm i\nu$  are the roots of the characteristic equation, while  $r, \delta$  are determined by the initial conditions, cf. (10.89). The second term — the solution to the homogeneous equation — is known as the *transient*, since it decays exponentially fast to zero. Thus, at large times, any internal motions of the system that might have been excited by the initial conditions die out, and only the particular solution (10.107) incited by the continued forcing persists.

## Exercises

10.6.1. Graph the following functions. Describe the fast oscillatory and beat frequencies:

- (a)  $\cos 8t - \cos 9t$ , (b)  $\cos 26t - \cos 24t$ , (c)  $\cos 10t + \cos 9.5t$ , (d)  $\cos 5t - \sin 5.2t$ .

10.6.2. Solve the following initial value problems: (a)  $\ddot{u} + 36u = \cos 3t$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 0$ .

- (b)  $\ddot{u} + 6\dot{u} + 9u = \cos t$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 1$ . (c)  $\ddot{u} + \dot{u} + 4u = \cos 2t$ ,  $u(0) = 1$ ,  $\dot{u}(0) = -1$ . (d)  $\ddot{u} + 9u = 3 \sin 3t$ ,  $u(0) = 1$ ,  $\dot{u}(0) = -1$ . (e)  $2\ddot{u} + 3\dot{u} + u = \cos \frac{1}{2}t$ ,  $u(0) = 3$ ,  $\dot{u}(0) = -2$ . (f)  $3\ddot{u} + 4\dot{u} + u = \cos t$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 0$ .

10.6.3. Solve the following initial value problems. In each case, graph the solution and explain what type of motion is represented. (a)  $\ddot{u} + 4\dot{u} + 40u = 125 \cos 5t$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 0$ ,

- (b)  $\ddot{u} + 25u = 3 \cos 4t$ ,  $u(0) = 1$ ,  $\dot{u}(0) = 1$ , (c)  $\ddot{u} + 16u = \sin 4t$ ,  $u(0) = 0$ ,  $\dot{u}(0) = 0$ , (d)  $\ddot{u} + 6\dot{u} + 5u = 25 \sin 5t$ ,  $u(0) = 4$ ,  $\dot{u}(0) = 2$ .

- 10.6.4. A mass  $m = 25$  is attached to a unit spring with  $k = 1$ , and frictional coefficient  $\beta = .01$ . The spring will break when it moves more than 1 unit. Ignoring the effect of the transient, what is the maximum allowable amplitude  $\alpha$  of periodic forcing at frequency  $\gamma =$
- .19?
  - .2?
  - .21?

- 10.6.5. (a) For what range of frequencies  $\gamma$  can you force the mass in Exercise 10.6.4 with amplitude  $\alpha = .5$  without breaking the spring? (b) How large should the friction be so that you can safely force the mass at any frequency?

- 10.6.6. Suppose the mass–spring–oil system of Exercise 10.5.39(b) is subject to a periodic external force  $2 \cos 2t$ . Discuss, in as much detail as you can, the long-term motion of the mass.

- ◇ 10.6.7. Write down the solution  $u(t, \gamma)$  to the initial value problem  $m \frac{d^2u}{dt^2} + k u = \alpha \cos \gamma t$ ,
- $u(0) = \dot{u}(0) = 0$ , for (a) a non-resonant forcing function at frequency  $\gamma \neq \omega$ ;
  - a resonant forcing function at frequency  $\gamma = \omega$ .
  - Show that, as  $\gamma \rightarrow \omega$ , the limit of the non-resonant solution equals the resonant solution. Conclude that the solution  $u(t, \gamma)$  depends continuously on the frequency  $\gamma$  even though its mathematical formula changes significantly at resonance.
- ◇ 10.6.8. Justify the solution formulas (10.107) and (10.108).
- ♣ 10.6.9. (a) Does a function of the form  $u(t) = a \cos \gamma t - b \cos \omega t$  still exhibit beats when  $\gamma \approx \omega$ , but  $a \neq b$ ? Use a computer to graph some particular cases and discuss what you observe.
- Explain to what extent the conclusions based on (10.103) do not depend upon the choice of initial conditions (10.102).

## Electrical Circuits

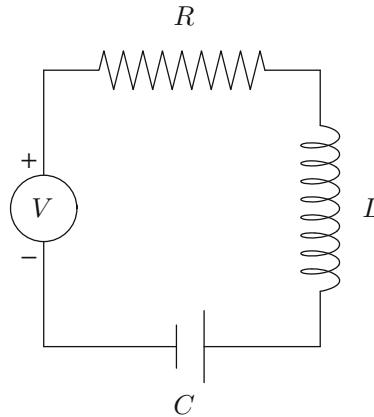
The Electrical–Mechanical Correspondence outlined in Section 6.2 will continue to operate in the dynamical universe. The equations governing the equilibria of simple electrical circuits and the mechanical systems such as mass–spring chains and structures all have the same underlying mathematical structure. In a similar manner, although they are based on a completely different set of physical principles, circuits with dynamical currents and voltages are modeled by second order linear dynamical systems of the Newtonian form presented earlier.

In this section, we briefly analyze the very simplest situation: a single loop containing a *resistor*  $R$ , an *inductor*  $L$ , and a *capacitor*  $C$ , as illustrated in Figure 10.14. This basic *RLC circuit* serves as the prototype for more general electrical networks linking various resistors, inductors, capacitors, batteries, voltage sources, etc. (Extending the mathematical analysis to more complicated circuits would make an excellent in-depth student research project.) Let  $u(t)$  denote the current in the circuit at time  $t$ . We use  $v_R, v_L, v_C$  to denote the induced voltages in the three circuit elements; these are prescribed by the fundamental laws of electrical circuitry.

- First, as we learned in Section 6.2, the resistance  $R \geq 0$  is the proportionality factor between voltage and current, so  $v_R = R u$ .
- The voltage passing through an inductor is proportional to the rate of change in the current. Thus,  $v_L = L \dot{u}$ , where  $L > 0$  is the *inductance*, and the dot indicates time derivative.
- On the other hand, the current passing through a capacitor is proportional to the rate of change in the voltage, and so  $u = C \dot{v}_C$ , where  $C > 0$  denotes the *capacitance*.

We integrate<sup>†</sup> this relation to produce the capacitor voltage  $v_C = \int \frac{u(t)}{C} dt$ .

<sup>†</sup> The integration constant is not important, since we will differentiate the resulting equation.



**Figure 10.14.** The Basic  $RLC$  Circuit.

The *Voltage Balance Law* tells us that the total of these individual voltages must equal any externally applied voltage  $v_E = F(t)$  coming from, say, a battery or generator. Therefore,

$$v_R + v_L + v_C = v_E.$$

Substituting the preceding formulas, we deduce that the current  $u(t)$  in our circuit satisfies the following linear integro-differential equation:

$$L \frac{du}{dt} + R u + \int \frac{u}{C} dt = F(t). \quad (10.110)$$

We can convert this into a differential equation by differentiating both sides with respect to  $t$ . Assuming, for simplicity, that  $L$ ,  $R$ , and  $C$  are constant, the result is the linear second order ordinary differential equation

$$L \frac{d^2u}{dt^2} + R \frac{du}{dt} + \frac{1}{C} u = f(t) = F'(t). \quad (10.111)$$

The current will be uniquely specified by the initial conditions  $u(t_0) = a$ ,  $\dot{u}(t_0) = b$ .

Comparing (10.111) with the equation (10.87) for a mechanically vibrating mass, we see that the correspondence between electrical circuits and mechanical structures developed in Chapter 6 continues to hold in the dynamical regime. The current  $u$  corresponds to the displacement. The inductance  $L$  plays the role of mass, the resistance  $R$  corresponds, as before, to friction, while the reciprocal  $1/C$  of capacitance is analogous to the spring stiffness. Thus, all of our analytical conclusions regarding stability of equilibria, qualitative behavior, solution formulas, etc., that we established in the mechanical context can, suitably re-interpreted, be immediately applied to electrical circuit theory.

In particular, an  $RLC$  circuit is *underdamped* if  $R^2 < 4L/C$ , and the current  $u(t)$  oscillates with frequency

$$\nu = \sqrt{\frac{1}{C L} - \frac{R^2}{4L^2}}, \quad (10.112)$$

while dying off to zero at an exponential rate  $e^{-Rt/(2L)}$ . In the overdamped and critically damped cases  $R^2 \geq 4L/C$ , the resistance in the circuit is so large that the current merely decays to zero at an exponential rate and no longer exhibits any oscillatory behavior. Attaching an alternating current source  $F(t) = \alpha \cos \gamma t$  to the circuit can induce a

catastrophic resonance if there is no resistance and the forcing frequency is equal to the circuit's natural frequency.

---

## Exercises

- 10.6.10. Classify the following  $RLC$  circuits as (i) underdamped, (ii) critically damped, or (iii) overdamped: (a)  $R = 1$ ,  $L = 2$ ,  $C = 4$ , (b)  $R = 4$ ,  $L = 3$ ,  $C = 1$ , (c)  $R = 2$ ,  $L = 3$ ,  $C = 3$ , (d)  $R = 4$ ,  $L = 10$ ,  $C = 2$ , (e)  $R = 1$ ,  $L = 1$ ,  $C = 3$ .
- 10.6.11. Find the current in each of the unforced  $RLC$  circuits in Exercise 10.6.10 induced by the initial data  $u(0) = 1$ ,  $\dot{u}(0) = 0$ .
- 10.6.12. A circuit with  $R = 1$ ,  $L = 2$ ,  $C = 4$ , includes an alternating current source  $F(t) = 25 \cos 2t$ . Find the solution to the initial value problem  $u(0) = 1$ ,  $\dot{u}(0) = 0$ .
- 10.6.13. A superconducting  $LC$  circuit has no resistance:  $R = 0$ . Discuss what happens when the circuit is wired to an alternating current source  $F(t) = \alpha \cos \gamma t$ .
- 10.6.14. A circuit with  $R = .002$ ,  $L = 12.5$ , and  $C = 50$  can carry a maximum current of 250. Ignoring the effect of the transient, what is the maximum allowable amplitude  $\alpha$  of an applied periodic current  $F(t) = \alpha \cos \gamma t$  at frequency  $\gamma =$  (a) .04? (b) .05? (c) .1?
- 10.6.15. Given the circuit in Exercise 10.6.14, over what range of frequencies  $\gamma$  can you supply a unit amplitude periodic current source?
- 10.6.16. How large should the resistance in the circuit in Exercise 10.6.14 be so that you can safely apply any unit amplitude periodic current?
- 

## Forcing and Resonance in Systems

Let us conclude by briefly discussing the effect of periodic forcing on a system of second order ordinary differential equations. Periodically forcing an undamped mass–spring chain or structure, or a resistanceless electrical network, leads to a second order system of the form

$$M \frac{d^2 \mathbf{u}}{dt^2} + K \mathbf{u} = \cos(\gamma t) \mathbf{a}. \quad (10.113)$$

Here  $M > 0$  and  $K \geq 0$  are  $n \times n$  matrices as above, cf. (10.82), while  $\mathbf{a} \in \mathbb{R}^n$  is a constant vector representing both a magnitude and a “direction” of the forcing and  $\gamma$  is the forcing frequency. Superposition is used to determine the effect of several such forcing functions. As always, the solution to the inhomogeneous system is composed of one particular response to the external force combined with the general solution to the homogeneous system, which, in the stable case  $K > 0$ , is a quasi-periodic combination of the normal vibrational modes.

To find a particular solution to the inhomogeneous system, let us try the trigonometric ansatz

$$\mathbf{u}^*(t) = \cos(\gamma t) \mathbf{w} \quad (10.114)$$

in which  $\mathbf{w}$  is a constant vector. Substituting into (10.113) leads to a linear algebraic system

$$(K - \mu M) \mathbf{w} = \mathbf{a}, \quad \text{where} \quad \mu = \gamma^2. \quad (10.115)$$

If the linear system (10.115) has a solution, then our ansatz (10.114) is valid, and we have produced a particular vibration of the system (10.113) possessing the same frequency as the forcing vibration. In particular, if  $\mu = \gamma^2$  is *not* a generalized eigenvalue of the matrix pair  $K, M$ , as described in (10.85), then the coefficient matrix  $K - \mu M$  is nonsingular, and

so (10.115) can be uniquely solved for any right-hand side  $\mathbf{a}$ . The general solution, then, will be a quasi-periodic combination of this particular solution coupled with the normal mode vibrations at the natural, unforced frequencies of the system.

The more interesting case occurs when  $\gamma^2 = \mu$  is a generalized eigenvalue, and so  $K - \mu M$  is singular, its kernel being equal to the generalized eigenspace  $V_\mu = \ker(K - \mu M)$ . In this case, (10.115) will have a solution  $\mathbf{w}$  if and only if  $\mathbf{a}$  lies in the image of  $K - \mu M$ . According to the Fredholm Alternative Theorem 4.46, the image is the orthogonal complement of the cokernel, which, since the coefficient matrix is symmetric, is the same as the kernel. Therefore, (10.115) will have a solution if and only if  $\mathbf{a}$  is orthogonal to  $V_\mu$ , i.e.,  $\mathbf{a} \cdot \mathbf{v} = \mathbf{0}$  for every generalized eigenvector  $\mathbf{v} \in V_\mu$ . Thus, one can force a system at a natural frequency without inciting resonance, provided that the “direction” of forcing, as determined by the vector  $\mathbf{a}$ , is orthogonal — in the linear algebraic sense — to the natural directions of motion of the system, as governed by the eigenvectors for that particular frequency.

If the orthogonality condition is not satisfied, then the periodic solution ansatz (10.114) does not apply, and we are in a truly resonant situation. Inspired by the scalar solution, let us try a *resonant solution ansatz*

$$\mathbf{u}^*(t) = t \sin(\gamma t) \mathbf{y} + \cos(\gamma t) \mathbf{w}. \quad (10.116)$$

Since

$$\frac{d^2 \mathbf{u}^*}{dt^2} = -\gamma^2 t \sin(\gamma t) \mathbf{y} + \cos(\gamma t) (2\gamma \mathbf{y} - \gamma^2 \mathbf{w}),$$

the function (10.116) will solve the differential equation (10.113) provided

$$(K - \mu M) \mathbf{y} = \mathbf{0}, \quad (K - \mu M) \mathbf{w} = \mathbf{a} - 2\gamma \mathbf{y}, \quad \mu = \gamma^2. \quad (10.117)$$

The first equation requires that  $\mathbf{y} \in V_\mu$  be a generalized eigenvector of the matrix pair  $K, M$ . Again, the Fredholm Alternative implies that, since the coefficient matrix  $K - \mu M$  is symmetric, the second equation will be solvable for  $\mathbf{w}$  if and only if  $\mathbf{a} - 2\gamma \mathbf{y}$  is orthogonal to the generalized eigenspace  $V_\mu = \text{coker}(K - \mu M) = \ker(K - \mu M)$ . Thus, the vector  $2\gamma \mathbf{y}$  is required to be the orthogonal projection of  $\mathbf{a}$  onto the eigenspace  $V_\mu$ . With this choice of  $\mathbf{y}$  and  $\mathbf{w}$ , formula (10.116) defines the resonant solution to the system.

**Theorem 10.43.** An undamped vibrational system will be periodically forced into resonance if and only if the forcing  $\mathbf{f} = \cos(\gamma t) \mathbf{a}$  is at a natural frequency of the system and the direction of forcing  $\mathbf{a}$  is not orthogonal to the natural direction(s) of motion of the system at that frequency.

**Example 10.44.** Consider the periodically forced system

$$\frac{d^2 \mathbf{u}}{dt^2} + \begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos t \\ 0 \end{pmatrix}.$$

The eigenvalues of the coefficient matrix are  $\lambda_1 = 5$ ,  $\lambda_2 = 1$ , with corresponding orthogonal eigenvectors  $\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ ,  $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ . The internal frequencies are  $\omega_1 = \sqrt{\lambda_1} = \sqrt{5}$ ,

---

<sup>†</sup> We can safely ignore the arbitrary multiple of the generalized eigenvector that can be added to  $\mathbf{w}$  as we only need find one particular solution; these will reappear anyway once we assemble the general solution to the system.

$\omega_2 = \sqrt{\lambda_2} = 1$ , and hence we are forcing at a resonant frequency. To obtain the resonant solution (10.116), we first note that  $\mathbf{a} = (1, 0)^T$  has orthogonal projection  $\mathbf{p} = \left(\frac{1}{2}, \frac{1}{2}\right)^T$  onto the eigenline spanned by  $\mathbf{v}_2$ , and hence  $\mathbf{y} = \frac{1}{2}\mathbf{p} = \left(\frac{1}{4}, \frac{1}{4}\right)^T$ . We can then solve

$$(K - I)\mathbf{w} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \mathbf{w} = \mathbf{a} - \mathbf{p} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \quad \text{for}^\dagger \quad \mathbf{w} = \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix}.$$

Therefore, the particular resonant solution is

$$\mathbf{u}^*(t) = (t \sin t) \mathbf{y} + (\cos t) \mathbf{w} = \begin{pmatrix} \frac{1}{4}t \sin t + \frac{1}{4} \cos t \\ \frac{1}{4}t \sin t \end{pmatrix}.$$

The general solution to the system is

$$\mathbf{u}(t) = \begin{pmatrix} \frac{1}{4}t \sin t + \frac{1}{4} \cos t \\ \frac{1}{4}t \sin t \end{pmatrix} + r_1 \cos(\sqrt{5}t - \delta_1) \begin{pmatrix} -1 \\ 1 \end{pmatrix} + r_2 \cos(t - \delta_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where the amplitudes  $r_1, r_2$  and phase shifts  $\delta_1, \delta_2$ , are fixed by the initial conditions. Eventually the resonant terms involving  $t \sin t$  dominate the solution, inducing progressively larger and larger oscillations.

## Exercises

10.6.17. Find the general solution to the following forced second order systems:

- (a)  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 7 & -2 \\ -2 & 4 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos t \\ 0 \end{pmatrix}$ , (b)  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 5 & -2 \\ -2 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 0 \\ 5 \sin 3t \end{pmatrix}$ ,  
 (c)  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 13 & -6 \\ -6 & 8 \end{pmatrix} \mathbf{u} = \begin{pmatrix} 5 \cos 2t \\ \cos 2t \end{pmatrix}$ , (d)  $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos \frac{1}{2}t \\ -\cos \frac{1}{2}t \end{pmatrix}$ ,  
 (e)  $\begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix} \frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos t \\ 11 \sin 2t \end{pmatrix}$ , (f)  $\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 6 & -4 & 1 \\ -4 & 6 & -1 \\ 1 & -1 & 11 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos t \\ 0 \\ \cos t \end{pmatrix}$ .

10.6.18. (a) Find the resonant frequencies of a mass–spring chain consisting of two masses,  $m_1 = 1$  and  $m_2 = 2$ , connected to top and bottom supports by identical springs with unit stiffness. (b) Write down an explicit forcing function that will excite the resonance.

10.6.19. Suppose one of the fixed supports is removed from the mass–spring chain of Exercise 10.6.18. Does your forcing function still excite the resonance? Do the internal vibrations of the masses (i) speed up, (ii) slow down, or (iii) remain the same? Does your answer depend upon which of the two supports is removed?

- ♣ 10.6.20. Find the resonant frequencies of the following structures, assuming the nodes all have unit mass. Then find a means of forcing the structure at one of the resonant frequencies, and yet not exciting the resonance. Can you also force the structure without exciting any mechanism or rigid motion? (a) The square truss of Exercise 6.3.5; (b) the joined square truss of Exercise 6.3.6; (c) the house of Exercise 6.3.8; (d) the triangular space station of Example 6.6; (e) the triatomic molecule of Example 10.41; (f) the water molecule of Exercise 10.5.30.

# References

- [1] Abraham, R., Marsden, J.E., and Ratiu, T., *Manifolds, Tensor Analysis, and Applications*, Springer–Verlag, New York, 1988.
- [2] Apostol, T.M., *Calculus*, Blaisdell Publishing Co., Waltham, Mass., 1967–69.
- [3] Baker, G.A., Jr., and Graves–Morris, P., *Padé Approximants*, Encyclopedia of Mathematics and Its Applications, v. 59, Cambridge Univ. Press, Cambridge, 1996.
- [4] Behrends, E., *Introduction to Markov Chains*, Vieweg, Braunschweig/Wiesbaden, Germany, 2000.
- [5] Bertalmío, M., *Image Processing for Cinema*, Cambridge University Press, Cambridge, 2013.
- [6] Bollobás, B., *Graph Theory: An Introductory Course*, Graduate Texts in Mathematics, vol. 63, Springer–Verlag, New York, 1993.
- [7] Boyce, W.E., and DiPrima, R.C., *Elementary Differential Equations and Boundary Value Problems*, Seventh Edition, John Wiley & Sons, Inc., New York, 2001.
- [8] Bradie, B., *A Friendly Introduction to Numerical Analysis*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2006.
- [9] Brigham, E.O., *The Fast Fourier Transform*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974.
- [10] Briggs, W.L., and Henson, V.E., *The DFT. An Owner’s Manual for the Discrete Fourier Transform*, SIAM, Philadelphia, PA, 1995.
- [11] Bürgisser, P., Clausen, M., and Shokrollahi, M.A., *Algebraic Complexity Theory*, Springer–Verlag, New York, 1997.
- [12] Buss, S.A., *3D Computer Graphics*, Cambridge University Press, Cambridge, 2003.
- [13] Cantwell, B.J., *Introduction to Symmetry Analysis*, Cambridge University Press, Cambridge, 2003.
- [14] Chung, F.R.K., *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics, No. 92, Amer. Math. Soc., Providence, R.I., 1997.
- [15] Cooley, J.W., and Tukey, J.W., An algorithm for the machine computation of complex Fourier series, *Math. Comp.* **19** (1965), 297–301.
- [16] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, Interscience Publ., New York, 1953.
- [17] Crowe, M.J., *A History of Vector Analysis*, Dover Publ., New York, 1985.
- [18] Daubechies, I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.
- [19] Davidson, K.R., and Donsig, A.P., *Real Analysis with Real Applications*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2002.
- [20] DeGroot, M.H., and Schervish, M.J., *Probability and Statistics*, Third Edition, Addison–Wesley, Boston, 2002.
- [21] Demmel, J.W., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
- [22] Diacu, F., *An Introduction to Differential Equations*, W.H. Freeman, New York, 2000.
- [23] Durrett, R., *Essentials of Stochastic Processes*, Springer–Verlag, New York, 1999.
- [24] Enders, C.K., *Applied Missing Data Analysis*, The Guilford Press, New York, 2010.

- [25] Farin, G.E., *Curves and Surfaces for CAGD: A Practical Guide*, Academic Press, London, 2002.
- [26] Fine, B., and Rosenberger, G., *The Fundamental Theorem of Algebra*, Undergraduate Texts in Mathematics, Springer–Verlag, New York, 1997.
- [27] Fortnow, L., *The Golden Ticket: P, NP, and the Search for the Impossible*, Princeton University Press, Princeton, N.J., 2013.
- [28] Foucart, S., and Rauhut, H., *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, Springer, New York, 2013.
- [29] Francis, J.G.F., The *QR* transformation I, II, *Comput. J.* **4** (1961–2), 265–271, 332–345.
- [30] Gohberg, I., and Koltracht, I., Triangular factors of Cauchy and Vandermonde matrices, *Integral Eq. Operator Theory* **26** (1996), 46–59.
- [31] Goldstein, H., *Classical Mechanics*, Second Edition, Addison–Wesley, Reading, MA, 1980.
- [32] Golub, G.H., and Van Loan, C.F., *Matrix Computations*, Third Edition, Johns Hopkins Univ. Press, Baltimore, 1996.
- [33] Graver, J.E., *Counting on Frameworks: Mathematics to Aid the Design of Rigid Structures*, Dolciani Math. Expo. No. 25, Mathematical Association of America, Washington, DC, 2001.
- [34] Guckenheimer, J., and Holmes, P., *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci., vol. 42, Springer–Verlag, New York, 1983.
- [35] Haar, A., Zur Theorie der orthogonalen Funktionensysteme, *Math. Ann.* **69** (1910), 331–371.
- [36] Hale, J.K., *Ordinary Differential Equations*, Second Edition, R. E. Krieger Pub. Co., Huntington, N.Y., 1980.
- [37] Herrlich, H., and Strecker, G.E., *Category Theory; an Introduction*, Allyn and Bacon, Boston, 1973.
- [38] Herstein, I.N., *Abstract Algebra*, John Wiley & Sons, Inc., New York, 1999.
- [39] Hestenes, M.R., and Stiefel, E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* **49** (1952), 409–436.
- [40] Higham, N.J., *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, Philadelphia, 2002.
- [41] Hirsch, M.W., and Smale, S., *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.
- [42] Hobson, E.W., *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, Dover Publ., New York, 1957.
- [43] Hogg, R.V., Tanis, E.A., and Zimmerman, D.L., *Probability and Statistical Inference*, Ninth Edition, Pearson Education Inc., Boston, MA, 2013.
- [44] Hoggatt, V.E., Jr., and Lind, D.A., The dying rabbit problem, *Fib. Quart.* **7** (1969), 482–487.
- [45] Hoory, S., Linial, N., and Wigderson, A., Expander Graphs and Their Applications, *Bull. Amer. Math. Soc.* **43** (2006), 439–561.
- [46] Hotelling, H., Analysis of complex of statistical variables into principal components, *J. Educ. Psychology* **24** (1933), 417–441, 498–520.
- [47] Jolliffe, I.T., *Principal Component Analysis*, Second Edition, Springer–Verlag, New York, 2002.

- [48] Kato, T., *Perturbation Theory for Linear Operators*, Corrected Printing of Second Edition, Springer–Verlag, New York, 1980.
- [49] Keener, J.P., *Principles of Applied Mathematics. Transformation and Approximation*, Addison–Wesley Publ. Co., New York, 1988.
- [50] Krall, A.M., *Applied Analysis*, D. Reidel Publishing Co., Boston, 1986.
- [51] Kublanovskaya, V.N., On some algorithms for the solution of the complete eigenvalue problem, *USSR Comput. Math. Math. Phys.* **3** (1961), 637–657.
- [52] Langville, A.N., and Meyer, C.D., *Google’s PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ, 2006..
- [53] Mandelbrot, B.B., *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.
- [54] Messiah, A., *Quantum Mechanics*, John Wiley & Sons, New York, 1976.
- [55] Misner, C.W., Thorne, K.S., and Wheeler, J.A., *Gravitation*, W.H. Freeman, San Francisco, 1973.
- [56] Moon, F.C., *Chaotic Vibrations*, John Wiley & Sons, New York, 1987.
- [57] Murray, R.N., Li, Z.X., and Sastry, S.S., *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.
- [58] Nilsson, J.W., and Riedel, S., *Electric Circuits*, Seventh Edition, Prentice–Hall, Inc., Upper Saddle River, N.J., 2005.
- [59] Olver, F.W.J., Lozier, D.W., Boisvert, R.F., and Clark, C.W., eds., *NIST Handbook of Mathematical Functions*, Cambridge University Press, Cambridge, 2010.
- [60] Olver, P.J., *Applications of Lie Groups to Differential Equations*, Second Edition, Graduate Texts in Mathematics, vol. 107, Springer–Verlag, New York, 1993.
- [61] Olver, P.J., *Introduction to Partial Differential Equations*, Undergraduate Texts in Mathematics, Springer, New York, 2014.
- [62] Ortega, J.M., *Numerical Analysis; a Second Course*, Academic Press, New York, 1972.
- [63] Oruç, H., and Phillips, G. M., Explicit factorization of the Vandermonde matrix, *Linear Algebra Appl.* **315** (2000), 113–123.
- [64] Page, L., Brin, S., Motwani, R., Winograd, T.; The PageRank citation ranking: bringing order to the web; Technical Report, Stanford University, 1998.
- [65] Pearson, K., On lines and planes of closest fit to systems of points in space, *Phil. Mag.* **2** (1901), 559–572.
- [66] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, Cambridge University Press, Cambridge, 1995.
- [67] Reed, M., and Simon, B., *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.
- [68] Royden, H.L., *Real Analysis*, Macmillan Co., New York, 1988.
- [69] Saad, Y., *Numerical Methods for Large Eigenvalue Problems*, Classics Appl. Math., vol. 66, SIAM, Philadelphia, 2011.
- [70] Saad, Y., *Iterative Methods for Sparse Linear Systems*, Second Edition, SIAM, Philadelphia, 2003.
- [71] Saad, Y., and Schultz, M.H., GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.* **7** (1986), 856–869.

- [72] Salomon, D., *Computer Graphics and Geometric Modeling*, Springer–Verlag, New York, 1999.
- [73] Sapiro, G., *Geometric Partial Differential Equations and Image Analysis*, Cambridge Univ. Press, Cambridge, 2001.
- [74] Schumaker, L.L., *Spline Functions: Basic Theory*, John Wiley & Sons, New York, 1981.
- [75] Sommese, A.J., and Wampler, C.W., *Numerical Solution of Polynomial Systems Arising in Engineering and Science*, World Scientific, Singapore, 2005.
- [76] Spielman, D., Spectral graph theory, in: *Combinatorial Scientific Computing*, U. Naumann and O. Schenk, eds., Chapman & Hall/CRC Computational Science, Boca Raton, Fl, 2012, pp. 495–524.
- [77] Stein, E.M., and Shakarchi, R., *Fourier Analysis: An Introduction*, Princeton Lectures in Analysis, Princeton University Press, Princeton, N.J., 2003.
- [78] Stewart, J., *Calculus: Early Transcendentals*, Fifth Edition, Thomson Brooks Cole, Belmont, CA, 2003.
- [79] Strang, G., *Introduction to Applied Mathematics*, Wellesley Cambridge Press, Wellesley, Mass., 1986.
- [80] Strang, G., *Linear Algebra and Its Applications*, Third Edition, Harcourt, Brace, Jovanovich, San Diego, 1988.
- [81] Strang, G., and Fix, G.J., *An Analysis of the Finite Element Method*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1973.
- [82] Strassen, V., Gaussian elimination is not optimal, *Numer. Math.* **13** (1969), 354–356.
- [83] Tannenbaum, P., *Excursions in Modern Mathematics*, Fifth Edition, Prentice–Hall, Inc., Upper Saddle River, N.J., 2004.
- [84] Tapia, R.A., Dennis, J.E., Jr., and Schäfermeyer, J.P., Inverse, shifted inverse, and Rayleigh quotient iteration as Newton’s method, *SIAM Rev.* **60** (2018), 3–55.
- [85] van der Vorst, H.A., *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press, Cambridge, 2003.
- [86] Varga, R.S., *Matrix Iterative Analysis*, Second Edition, Springer–Verlag, New York, 2000.
- [87] Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K., *Probability and Statistics for Scientists and Engineers*, Ninth Edition, Prentice–Hall, Inc., Upper Saddle River, N.J., 2012.
- [88] Walter, G.G., and Shen, X., *Wavelets and Other Orthogonal Systems*, Second Edition, Chapman & Hall/CRC, Boca Raton, Fl, 2001.
- [89] Watkins, D.S., *Fundamentals of Matrix Computations*, Wiley–Interscience, New York, 2002.
- [90] Wilkinson, J.H., *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
- [91] Wilson, E.B., Jr., Decius, J.C., and Cross, P.C., *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*, Dover Publ., New York, 1980.
- [92] Yaglom, I.M., *Felix Klein and Sophus Lie*, Birkhäuser, Boston, 1988.
- [93] Yale, P.B., *Geometry and Symmetry*, Holden–Day, San Francisco, 1968.

# Symbol Index

Symbol	Meaning	Page(s)
$c + d$	addition of scalars	8
$A + B$	addition of matrices	5
$\mathbf{v} + \mathbf{w}$	addition of vectors	5, 76
$V + W$	addition of subspaces	86
$f + g$	addition of functions	79
$cd$	multiplication of scalars	8
$c\mathbf{v}, cA, cf$	scalar multiplication	5, 76, 79
$AB, A\mathbf{x}$	matrix and vector multiplication	6
$K > 0$	positive definite	157, 398
$K \geq 0$	positive semi-definite	159
$ \cdot $	absolute value, modulus, norm	xvii, 137, 174
$\ \cdot\ $	norm	131, 133, 144, 178–9
$\ \cdot\ $	matrix norm	153
$\ \cdot\ _1$	1 norm	145, 155
$\ \cdot\ _2$	Euclidean norm	145, 460
$\ \cdot\ _\infty$	max norm	145, 155
$\ \cdot\ _p$	$p$ norm	145
$\ \cdot\ _{p,w}$	weighted $p$ norm	147
$\ \cdot\ _F$	Frobenius norm	156
$\mathbf{v} \cdot \mathbf{w}$	dot product	130
$\mathbf{z} \cdot \mathbf{w}$	Hermitian dot product	178
$\langle \cdot, \cdot \rangle$	inner product	130, 133, 179
$\langle\!\langle \cdot, \cdot \rangle\!\rangle$	inner product	133
$\langle\!\langle\!\langle \cdot, \cdot \rangle\!\rangle\!\rangle$	inner product	133
$[\cdot, \cdot]$	commutator	10, 354
$[a, b]$	closed interval	xvii, 79
$(a, b)$	open interval	79
$[a, b), (a, b]$	half open interval	79
$\{x \mid C\}$	set of all $x$ subject to conditions $C$	xvii
$x \simeq y$	approximate equality of real numbers	55
$V \simeq W$	isomorphic vector spaces	356
$V / W$	quotient vector space	57

$\#$	cardinality (number of elements)	<i>xvii</i>
$\cup$	union	<i>xvii</i>
$\cap$	intersection	<i>xvii</i>
$\in$	element of	<i>xvii</i>
$\notin$	not element of	<i>xvii</i>
$\subset, \subseteq, \subsetneq$	subset	<i>xvii</i>
$\supset$	superset	<i>xvii</i>
$\setminus$	set theoretic difference	<i>xvii</i>
$\binom{n}{k}$	binomial coefficient	58, 393
$\mathbf{u}^\perp$	orthogonal complement to vector	208
$W^\perp$	orthogonal complement	217
$\hat{x}$	angle	137
$f: X \rightarrow Y$	function	<i>xvii</i>
$x_n \rightarrow x$	convergent sequence	<i>xvii</i>
$\equiv$	equality of functions	<i>xvii</i>
$\circ$	composition	<i>xvii, 352</i>
$\times$	Cartesian product	81
$\times$	cross product	140, 187
$\sim$	agreement at sample points	287
$\sim_W$	equivalence relation	87
$[\cdot]_W$	equivalence class	87
$\bar{x}, \bar{f}$	mean or average	84, 467
$\bar{z}, \bar{\mathbf{z}}$	complex conjugate	173, 177, 390
$\bar{q}$	quaternion conjugate	365
$\sum_{i=1}^n$	summation	<i>xvii</i>
$\prod_{i=1}^n$	product	<i>xvii</i>
$u', u'', \dots$	space derivatives	<i>xviii, 173</i>
$\dot{u}, \ddot{u}, \dots$	time derivatives	<i>xviii, 600</i>
$\frac{du}{dx}, \frac{d^2u}{dx^2}, \dots$	ordinary derivatives	<i>xviii, 585</i>
$\frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x \partial t}, \dots$	partial derivatives	<i>xviii</i>
$\partial_x$	partial derivative operator	349
$\nabla$	gradient	349
$\nabla \cdot$	divergence	86
$\int f(x) dx$	indefinite integral	<i>xviii</i>
$\int_a^b f(x) dx$	definite integral	<i>xviii</i>

<b>0</b>	zero vector	7
$A^{-1}, L^{-1}$	inverse matrix, linear function	31, 355
$A^T$	transpose matrix	43, 502
$A^{-T}$	inverse transpose matrix	44
$A^\dagger$	Hermitian transpose matrix	181, 444
$A^+$	pseudoinverse matrix	437
$A^*, L^*$	adjoint matrix, linear function	395, 399
$L^*$	dual linear function	369
$V^*$	dual vector space	350
$\mathcal{A}$	space of analytic functions	84
$(A   \mathbf{b})$	augmented matrix	12
$a_{ij}$	matrix entry	<i>xvii</i> , 4
$B_1$	unit ball	148
$\mathbb{C}$	complex numbers	<i>xvii</i> , 173
$\mathbb{C}^n$	$n$ -dimensional complex space	177
$C_j$	cardinal spline	284
$C^0$	space of continuous functions	83, 179, 347
$C^n$	space of continuously differentiable functions	84
$C^\infty$	space of smooth functions	84
coker	cokernel	113, 357
coimg	coimage	113, 357
cos	cosine function	<i>xvii</i> , 176
cosh	hyperbolic cosine function	176
$D$	derivative operator	348
$d(\mathbf{v}, \mathbf{w})$	distance	146
$D_A$	Gerschgorin domain	420
$D_A^*$	refined Gerschgorin domain	422
$D_k$	Gerschgorin disk	420
$D^k$	$k^{\text{th}}$ derivative operator	353
$\mathcal{D}^{(k)}$	space of $k^{\text{th}}$ order differential operators	355
det	determinant	32, 70
diag	diagonal matrix	8
dim	dimension	100
$e$	base of natural logarithm	<i>xvii</i>
$e^x$	exponential function	<i>xvii</i>
$e^z$	complex exponential	175
$e^A$	matrix exponential	593
$\mathbf{e}_i$	standard basis vector	36, 99
$E_i$	elementary matrix	16
$\mathcal{F}$	function space	79, 80
$G_n$	complete graph	127, 464

$G_{m,n}$	complete bipartite digraph	127
$\mathbb{H}$	quaternions	364
$H^1$	Sobolev norm	136
$H_n$	Hilbert matrix	57
$i = \sqrt{-1}$	imaginary unit	<i>xvii</i> , 173
$i$	unit quaternion	364
$\mathbf{i}$	standard basis vector in $\mathbb{R}^3$	99
$\mathbf{I}, \mathbf{I}_n$	identity matrix	7
$\mathbf{I}, \mathbf{I}_V$	identity function	343
$\text{img}$	image	105, 383
$\text{Im}$	imaginary part	173, 177, 391
$\text{index}$	index of linear map	357
$j$	unit quaternion	364
$\mathbf{j}$	standard basis vector in $\mathbb{R}^3$	99
$J_{a,n}, J_{\lambda,n}$	Jordan block matrix	416, 449
$k$	unit quaternion	364
$\mathbf{k}$	standard basis vector in $\mathbb{R}^3$	99
$\ker$	kernel	105, 378
$L^2$	Hilbert space, inner product, norm	133, 180
$L^p$	function space $p$ norm	145
$L_k$	Lagrange polynomial	262
$L[\mathbf{v}]$	linear function	342
$\mathcal{L}(V, W)$	space of linear functions	349
$\log$	natural (base $e$ ) logarithm	<i>xvii</i>
$\log_a$	base $a$ logarithm	<i>xvii</i>
$m_A$	minimal polynomial	453
$\max$	maximum	<i>xvii</i>
$\min$	minimum	<i>xvii</i>
$\mathcal{M}_{m \times n}$	space of $m \times n$ matrices	77, 349
$\text{mod}$	modular arithmetic	<i>xvii</i>
$\mathbf{O}, \mathbf{O}_{m \times n}$	zero matrix	7
$\mathcal{O}^+$	nonnegative orthant	83
$p_A$	characteristic polynomial	415
$P_k$	Legendre polynomial	228
$\mathcal{P}^{(n)}$	space of polynomials of degree $\leq n$	77
$\mathcal{P}^{(\infty)}$	space of all polynomials	83
$\text{ph}$	phase	<i>xvii</i> , 174
$\mathbb{Q}$	rational numbers	<i>xvii</i>
$\mathbb{R}$	real numbers	<i>xvii</i>
$\mathbb{R}^n$	$n$ -dimensional real Euclidean space	76
$\mathbb{R}^\infty$	space of infinite sequences	81

$R_\theta$	rotation	349
rank	rank	61
$\text{Re}$	real part	173, 177, 391
$S_1$	unit sphere	83, 148
$S_j$	sinc function	273
sec	secant function	<i>xvii</i>
sech	hyperbolic secant function	270
sign	sign of permutation	72
sin	sine function	<i>xvii</i> , 176
sinh	hyperbolic sine function	176
span	span	87
supp	support of a function	551
$\mathbf{t}_k$	monomial sample vector	265
$T_k$	Chebyshev polynomial	233
$\mathcal{T}^{(n)}$	space of trigonometric polynomials of degree $\leq n$	90, 190
$\mathcal{T}^{(\infty)}$	space of all trigonometric polynomials	91
tan	tangent function	<i>xvii</i>
tr	trace	10, 415
$V^{(k)}$	Krylov subspace	537
$\mathbb{Z}$	integers	<i>xvii</i>
$\beta$	$B$ –spline	284
$\Delta$	Laplacian	349, 381
$\varepsilon_j$	dual basis vectors	350
$\zeta_n$	primitive root of unity	288
$\kappa$	condition number	460
$\lambda$	eigenvalue	408
$\pi$	area of unit circle	<i>xvii</i>
$\pi$	permutation	26, 27
$\rho$	spectral radius	489
$\rho_{xy}$	correlation	469
$\sigma_i$	singular value	454
$\sigma_x$	standard deviation	468
$\sigma_{xy}$	covariance	469
$\omega$	relaxation parameter	518
$\omega_k$	sampled exponential	287

# Subject Index

## A

absolute row sum 155, 496, 498, 510  
absolute value *xvii*  
abstract *vii*, *viii*, *viii*, *xi*, *xiv*, 6, 75–6  
acceleration 260, 608, 623  
account 476–7  
acute 160  
adapted basis 426  
addition 5, 48–9, 53, 76, 87, 161, 349, 576  
    associativity of 78  
    complex 173, 296, 298  
    matrix 5, 8, 12, 43, 349  
    real 296  
    quaternion 364  
    vector 77, 82, 390  
addition formula 89  
additive identity 76  
additive inverse 8, 76  
additive unit 7  
adjacency matrix 317  
adjoint *xii*, 112, 342, 357, 395–7, 399, 413  
    formal 396  
    Hermitian 181, 205  
    weighted 397  
adjoint function 398  
adjoint system 112, 117  
adjudicate 112  
advertising 258  
affine equation 479  
affine function *xi*, 245, 239, 343, 370, 555  
affine iterative system 488  
affine matrix 372, 603  
affine subspace 87, 375, 383  
affine system 488  
affine transformation *ix*, *xii*, *xiv*, 341, 370–3,  
    377, 419, 603  
air 259, 626  
airplane 200  
Albert Bridge 626  
algebra *viii*, 7, 98  
    abstract 7  
    computer 57  
    Fundamental Theorem of 98, 124, 415  
    Fundamental Theorem of Linear 114, 461  
    linear *vii*, *xi*, *xiii*, *xv*, 1, 75, 114, 126, 183,  
        243, 341, 403, 506  
    matrix 7, 99

algebra (*continued*)  
    numerical linear 48  
    polynomial 78  
algebraic function 166  
algebraic multiplicity 424  
algebraic system *vii*, *ix*, 341–2, 376, 386, 506,  
    517, 540  
algorithm *viii*, *ix*  
    numerical *viii*–*xi*, 48, 129, 183, 199, 400,  
        403, 475, 536, 547  
    QR *xii*, 200, 475, 527, 529, 531–2, 535–6,  
        538  
    tridiagonal solution 52, 282  
aliasing 286–7, 291  
alternating current 320, 629  
alternative  
    Fredholm *vii*, *ix*, *xi*, 183, 222, 226, 312,  
        330, 352, 377, 631  
altitude 269  
AM radio 293  
amplitude 89, 273, 565, 587, 609, 615, 627  
    phase- 89, 587, 610  
analysis *xv*, 129, 135  
    complex 381  
data *vii*, *viii*, *xii*, *xiii*, *xv*, 80, 126, 129, 135,  
    301, 403, 463  
discrete Fourier *xi*, *xiv*, *xv*, 235, 285  
Fourier *ix*, *viii*, *xii*, *xv*, 75, 78, 99, 135, 173,  
    180, 183, 188, 227, 285, 287, 476, 555  
functional *xii*  
linear *ix*, *x*  
numerical *vii*, *ix*, *xii*, *xiii*, *xv*, 1, 75, 78, 132,  
    156, 227, 230, 233, 235, 271, 279, 317,  
    475  
Principal Component *ix*, *xii*, *xiv*, *xv*, 255,  
    403, 467, 471–2  
real 151  
statistical 188, 238  
symmetry 599  
analytic 84, 87, 91  
angle *xvii*, 120, 129, 137–140, 149, 187, 419,  
    439, 525, 544, 545, 600  
acute 160  
Euler 203  
polar 174  
right 140, 184  
weighted 138

- animation *viii, xii, 200, 203, 279, 283, 341, 374, 375, 565, 599*  
     481
- ansatz *379–80*  
     exponential *379, 390, 567, 576, 621*  
     power *380, 479*  
     resonant *631*  
     trigonometric *609, 618, 623, 625, 630*
- anti-correlated *469*
- anti-derivative *xviii*
- anti-diagonal *86, 104*
- apparatus *565, 608*
- application *viii, 75*  
     applied mathematics *vii, ix, x, 1, 48, 230, 475*  
     approximate solution *237, 541*  
     approximation *viii, ix, xi, xiv, 220, 227, 261, 475, 542, 599*  
         dyadic *563*  
         Krylov *541–2*  
         least squares *188, 263, 272*  
         linear *324, 329, 341, 388*  
         matrix *462*  
         numerical *220, 235, 403, 416, 467*  
         Padé *261*  
         polynomial *266, 279*  
         rank *k* *462*  
         tangent line *600*  
         trigonometric *271, 273*
- architecture *483*
- area *361, 487*
- argument *xviii, 174*
- arithmetic *173, 413*  
     complex *173*  
     computer *57*  
     floating point *48, 58*  
     matrix *xiii, 1, 8, 77*  
     modular *xvii*  
     rational *58*  
     real *173*  
     single precision *461*
- arithmetic mean *148*
- arithmetic operation *48, 199, 212, 534, 548*
- Arnoldi matrix *540*
- Arnoldi Method *xii, 475, 538*
- Arnoldi vector *538–40, 542, 547*
- array *56*
- arrow *xvii, 568*
- art *375, 483*
- artifact  
     numerical *206*
- associative *5, 7–8, 76, 78, 365*
- astronomy *407*
- asymptotically stable *405, 478, 490, 493, 579–82, 584, 586–7, 591, 597*
- globally *405, 488–90, 492–3, 579, 622*
- Atlanta *504*
- atom *vii, 203, 437, 619, 620*
- audible frequency *287, 614*
- audio *102, 285, 293*
- augmented matrix *12, 24, 36, 60, 66–7*
- autonomous system *403, 566, 579*
- autonomous vehicle *616*
- average *10, 84, 256, 272, 288, 348, 467*
- axis *373, 419, 600*  
     coordinate *362*  
     imaginary *580–1*  
     principal *465, 472, 487*  
     semi- *487, 498*
- B**
- B*-spline *284, 567*
- Back Substitution *x, xiii, xiv, 3, 14, 21, 24, 41, 50, 53, 62, 208, 211, 282, 518*
- bacteria *406*
- balance *477–8, 499*  
     force *304, 327, 333*  
     voltage *312, 314, 629*
- ball *236, 244*  
     unit *85, 149–50, 473*
- banded matrix *55*
- bandwidth *55*
- bank *476, 478–9*
- bar *301, 320, 322–3, 328, 608*
- base *xvii*  
     data *555*
- basic variable *62–3, 118*
- basis *ix, x, xiii, 75, 99, 100–1, 177, 341, 343, 365, 403, 575, 577, 594*  
     adapted *426*  
     change of *365, 367*  
     dual *350, 352, 369*  
     eigenvector *xii, xvi, 183, 423, 427, 432, 434, 438, 446, 448, 480, 523, 528, 566, 572, 618*  
     Jordan *448, 450–1, 453, 480, 488, 576–7*  
     left-handed *103, 202, 222*  
     orthogonal *xi, xiii, xv, 184, 189, 194, 201, 214, 235, 266, 403, 435, 446, 551, 611, 618*  
     orthonormal *184, 188, 194–6, 198–9, 201, 204, 213, 235, 248, 288, 432, 437–8, 444, 456–7, 460, 475, 528–9, 538*  
     real *575*  
     right-handed *103, 201–2, 222*  
     standard *36, 99, 111, 184, 261, 343, 349, 356, 426, 449, 450, 529*  
     standard dual *350*  
     wavelet *102, 189, 204, 283, 550, 552, 555–6, 562*
- basis function *549, 562*

- battery 301, 311, 317–8, 320, 626, 629  
 beam 120, 279, 301, 322  
 beat 624, 628  
 bell-shaped 284  
 bend 322  
 benzene 620  
 bidiagonal 52–3, 402  
 bilinear 10, 130, 133, 156  
 bilinear form 156  
 bilinear function 347, 354  
 binary 297, 561, 563  
 Binet formula 483, 485  
 binomial coefficient 58–9, 380, 393  
 binomial formula 176, 393  
 biology *ix*, 1, 403, 407, 475, 499  
 bipartite 127  
 bit reversal 297  
 block  
     Jordan 416–7, 449–50, 453, 598  
 block diagonal 35, 74, 128, 171, 420, 449, 535, 598  
 block matrix 11, 35, 603  
 block upper triangular 74, 535  
 blue collar 504  
 body 200, 203, 259, 301, 341, 439  
 bolt 323  
 bond 620  
 Boston 504  
 boundary condition 302  
     clamped 280, 283–4  
     natural 280, 283–4  
     periodic 280, 283–4  
 boundary point 503  
 boundary value problem *x*, *xi*, *xv*, 54, 75, 92, 99, 136, 183, 222, 235, 322, 389, 397, 399, 541–2  
     linear *vii*, 342, 376–7, 386  
 bounded 219, 349, 380, 603  
 bounded support 557  
 bowl 236  
 box function 549, 555, 559  
 brain 287  
 bridge 625–6  
 Britain 625  
 bug 505  
 building 120, 322, 324  
 business 504
- C**
- C++ 14  
 CAD 279, 283  
 calculator 48, 260  
 calculus *vii*, *x*, 83, 231, 240, 341, 580  
     Fundamental Theorem of 347, 356, 606  
 calculus (*continued*)  
     multivariable *x*, 235, 242–3, 342, 441, 545, 582  
     vector 353, 365  
 calculus of variations *xii*, 235  
 canonical form 368  
     Jordan *xii*, *xiii*, 403, 447, 450, 490, 525, 598  
 capacitance 626  
 capacitor 311, 628  
 car 196, 254, 467  
 carbon 406, 434  
 carbon tetrachloride 620  
 cardinal spline 284  
 cardinality *xvii*  
 Cartesian coordinate 101  
 Cartesian product 81, 86, 133, 347, 377  
 category theory *viii*  
 Cauchy-Schwarz inequality 129, 137, 142–3, 179, 469  
 Cayley transform 204  
 Cayley–Hamilton Theorem 420, 453  
 Cayley–Klein parameter 203  
 CD 183, 283, 287  
 ceiling 322  
 center 373, 588, 589–91  
 center of mass 439  
 center manifold 605  
 Center Manifold Theorem 604  
 center subspace 604  
 CG — *see* Conjugate Gradient  
 chain  
     Jordan 447–8, 450–2, 488, 576–7, 579, 581, 603–4  
     Markov *xii*, 475, 499–502  
     mass–spring *xi*, *xiv*, 301, 309, 317, 399, 403, 565, 608, 610, 619, 628, 630  
     null Jordan 447, 451  
 chain rule 301  
 change of basis 365, 367  
 change of variables 172, 232, 234  
 chaos 611  
 characteristic equation 379–80, 390, 408–9, 413, 420, 453, 475, 567, 583, 586, 621, 627  
     generalized 435, 618  
 characteristic polynomial 408, 415, 453, 475  
 characteristic value 408  
 characteristic vector 408  
 charge 313  
 Chebyshev polynomial 233  
 chemistry 48, 139, 203, 403, 407, 608  
 Chicago 504  
 chlorine 620  
 Cholesky factorization 171

- circle 176, 329, 339, 363, 371, 438  
 unit *xvii*, 132, 288, 442, 530
- circuit *xiii*, 121–2, 124–6, 126, 312, 403, 608  
 electrical *viii*, *xii*, *xiv*, 122, 129, 129, 196, 235–6, 301, 628  
 fault-tolerant 464  
 LC 630  
 RLC 626, 629
- circuit vector 312
- circulant matrix 282, 436
- circular motion 616
- city 504–5
- clamped 280, 283–4
- clarinet 626
- class  
 equivalence 87
- classical mechanics 341, 388, 583
- climate 406
- clockwise 360
- closed *xvii*, 79, 146, 151
- closed curve 280, 283
- closest point *xi*, 183, 235, 238, 245–6, 298
- closure 82, 106
- cloudy 499, 501
- CMKY 470
- code  
 error correcting 464
- codomain *xvi*, *xvii*, 105, 342, 376, 383, 396, 618
- co-eigenvector 416, 503, 525
- coefficient  
 binomial 58–9, 380, 393  
 constant *x*, 363, 376–7, 390  
 Fourier 289, 291, 294, 296–7, 470  
 frictional 499, 504  
 leading 367  
 least squares 266  
 undetermined 372, 385–6, 500, 623  
 wavelet 470
- coefficient function 353
- coefficient matrix 4, 6, 63, 157, 224, 235, 241, 343, 476, 479, 484, 499, 508, 528, 531, 566, 575, 591, 606, 608, 618, 630
- cofactor 112
- coffee 486
- coimage *x*, 75, 113–5, 117, 221, 223–4, 357, 434, 457
- cokernel *x*, 75, 113–4, 116, 118, 125, 221–2, 312, 357, 434, 461, 470, 501, 542, 626, 631
- collapse 626
- collocation 547
- colony 406
- color 463, 470
- column 4, 6, 7, 27, 43, 45, 94, 114, 162, 201  
 pivot 56  
 orthonormal 444, 455–6  
 zero 59
- column interchange 57
- column operation 72, 74
- column permutation 418
- column space 105, 383
- column sum 10, 419, 501, 502
- column vector 4, 6, 46, 48, 7, 130, 350–17
- combination  
 linear 87, 95, 101, 287, 342, 388, 599, 618
- communication network 464
- commutation relation 355
- commutative 5, 6, 8, 76, 173, 352, 360
- commutator 10, 354, 601
- commuting matrix 10, 601
- compact 149
- company 258
- compatibility condition 96, 106, 183, 222
- compatible *ix*, *xi*, 8, 11, 62, 95, 224
- complement  
 orthogonal 217–9, 221, 431, 631
- complementary dimension 218
- complementary subspace 86, 105, 217–8, 221
- complete bipartite digraph 127
- complete eigenvalue *xvi*, 412, 424, 493, 531, 577, 588
- complete graph 127, 464
- complete matrix *xvi*, 403, 424–6, 428, 430–2, 444, 450, 480, 484, 490, 493, 522, 566, 572, 575, 603
- complete monomial polynomial 268
- complete the square *xi*, 129, 166, 240, 437
- completeness 562
- complex addition 173, 296, 298
- complex analysis 381
- complex arithmetic 173
- complex conjugate 173, 177, 205, 390, 444, 452
- complex diagonalizable 427
- complex eigenvalue *xiv*, 412, 415–6, 421, 423, 430, 433, 496, 525, 535, 538, 578, 587
- complex eigenvector *ix*, *xii*, *xiii*, 403, 408–10, 429, 443, 446, 448, 473, 475, 480, 484, 503, 522, 525, 527, 537, 539, 560, 565, 609
- complex exponential 175, 180, 183, 192, 285, 287, 390, 549
- complex inequality 177
- complex inner product 184
- complex iterative equation 478
- complex linear function 342
- complex linear system *xiv*, 566
- complex matrix 5, 181, 212, 226, 536, 566
- complex monomial 393

- complex multiplication 173, 296, 298  
complex number *xvii*, 80, 129, 173  
complex plane 173, 420, 580  
complex root 114, 390, 621  
complex scalar 177, 476  
complex solution *xiv*, 391, 575  
complex subspace 298, 430, 452  
complex trigonometric function 176–7  
complex-valued function 129, 173, 177, 179, 391  
complex variable 172  
complex vector 129, 433  
complex vector space *xi*, *xiv*, 76, 129, 177, 179, 287, 342, 390  
component  
    connected 124, 463  
    principal *ix*, *xii*, *xiv*, *xv*, 255, 403, 467, 471–2  
composition 79, 352  
compound interest 476, 478–9  
compressed sensing 238  
compression 99, 102, 183, 272, 293–4, 462, 552, 554–6, 558, 561–2  
computer *vii*, *xvi*, 48, 56–7, 260–1, 291, 461, 513, 561  
    binary 297  
    parallel 513  
    serial 513  
computer aided design 279, 283  
computer algebra 57  
computer arithmetic 57  
computer game *xii*, 200, 341, 375  
computer graphics *viii*, *xi*, 52, 200, 203, 279, 283, 286, 341, 358, 374–5, 565, 599  
computer programming 14, 28  
computer science *vii*, *ix*, *xv*, 126, 463  
computer software *xvi*, 404  
computer vision 375, 499  
condition  
    boundary 280, 283–4, 302  
    compatibility 96, 106, 183, 222  
    initial 404–5, 480, 570–2, 593, 609–10  
condition number 57, 460, 466  
    spectral 525, 591  
conditioned  
    ill- 56–7, 211, 249, 276–7, 461–2, 525  
conductance 313, 314, 317, 320  
cone 159–60  
conjugate  
    complex 173, 177, 205, 390, 444, 452  
    quaternion 365  
conjugate direction 543, 545, 548  
Conjugate Gradient 476, 542, 544–5, 548  
conjugate symmetry 179, 184  
conjugated 390–1  
connected 121, 124, 144, 463  
conservation of energy 585  
constant 348, 389  
    gravitational 259  
    integration 404  
    piecewise 551  
constant coefficient *x*, 363, 376–7, 390  
constant function 78  
constant polynomial 78  
constant solution 615  
constant vector 588  
constitutive equation 303  
constitutive relation 302, 312, 327  
constrained maximization principle 442–3  
constrained optimization principle 441  
continuous dynamics *xii*, 565  
continuous function *xi*, 83, 133, 150, 179, 219–20, 274, 347, 349, 377, 559  
continuous medium *ix*, 565  
continuously differentiable 84, 136, 379  
continuum mechanics *xi*, 235–6, 351, 399  
continuum physics 156  
contraction 600  
control system *vii*, *xv*, 76, 99, 106, 376  
control theory 235  
convergence *xv*, *xvii*, 91, 146, 151, 394, 475, 489, 506, 510, 514, 518, 545, 559  
    uniform 562  
convergence in norm 151  
convergent 91, 146, 151, 394  
convergent matrix 488–9, 495–6, 508, 517  
convergent sequence 86, 551  
convex 150  
cookbook *viii*  
coordinate 101, 151, 188–9, 350, 472  
    Cartesian 101  
    polar 90, 136, 174  
    principal 472, 477  
    radial 383  
coordinate axis 362  
coordinate plane 362  
coplanar 88  
core *xiii*  
corporation 504  
correlation 469–70  
    cross- 252  
correspondence  
    Electrical–Mechanical 321, 628  
cosine *xvii*, 138, 175–6, 183, 285, 581, 609–10  
    law of 139  
counterclockwise 359, 600  
country 504  
covariance 469, 470, 472  
covariance matrix 163, 470–1, 473  
Cramer's Rule 74

- criterion  
stopping 544
- critical point 240, 242
- critically damped 622, 629
- cross polytope 149
- cross product 140, 187, 239, 305, 602
- cross-correlation 252
- crystallography *xii*, 200, 358
- cube 126, 318
- cubic 260, 267, 280, 563  
piecewise 263, 269, 279–80
- curl 349
- current 122, 235, 311–3, 315, 318–20, 626–7  
alternating 320, 629
- Current Law 313–4, 317
- current source 301, 313–4, 317–8, 320, 629
- curve 86, 279, 568  
closed 280, 283  
level 585  
smooth 279  
space 283
- curve fitting 283
- cylinder 85
- D**
- damped 565, 498, 621, 623  
critically 622, 629
- damping *viii*, *ix*, *xii*, 302, 620
- data *xiv*, 235, 252, 467, 471  
experimental 237–8  
high-dimensional 471  
Lagrange 284  
missing 471  
normally distributed 473
- data analysis *vii*, *viii*, *xii*, *xiii*, *xv*, 80, 126, 129, 135, 301, 403, 463
- data base 555
- data compression 99, 183, 294, 462
- data fitting *xiii*, 132, 237, 254
- data matrix 462, 467, 470–1, 473  
normalized 470, 473
- data mining 467
- data point *xi*, 235, 237, 254–5, 272, 283, 467, 470, 474  
normalized 470–1
- data science 1, 235
- data set 188, 462, 472–3
- data transmission 272
- data vector 254
- Daubechies equation 558, 561
- Daubechies function 559–60
- Daubechies wavelet 555, 562
- daughter wavelet 550, 556, 563
- day 475, 477
- decay 627, 629  
exponential 405, 565, 580–1, 621  
radioactive 257, 404, 406
- decay rate 257, 404, 622
- decimal 561
- decomposition  
polar 439
- Schur *xii*, *xiii*, 403, 444–6
- spectral 440, 598
- Singular Value *xii*, 403, 455, 457, 461, 473
- decrease  
steepest 545, 583
- deer 406, 479
- definite 583  
negative 159–60, 171, 581, 583  
positive *xi*, *xiii*, *xiv*, 129, 156, 159–61, 164–5, 167, 170–1, 181–2, 204, 235, 241–2, 244, 246, 252, 301, 304, 309, 313, 316, 327, 396, 398, 432, 439, 443, 473, 528, 531, 542, 544, 581, 583, 608, 618, 622
- definite integral *xviii*
- deflation 420, 526
- deformation 438–9
- degree 78, 98, 161, 266, 273, 453, 537
- degree matrix 317
- denoising 102, 183, 293, 555, 562
- dense 220
- Department of Natural Resources 479
- dependent  
linearly 93, 95–6, 100, 571
- deposit 476, 479
- derivative *x*, *xviii*, 84, 177, 476, 542, 594  
ordinary *xviii*  
partial *xviii*, 242, 349
- derivative operator 348, 353
- descent 545, 549
- design *xi*, 235, 483  
font 203
- determinant *x*, *xiii*, *xiv*, 1, 32, 70, 72, 103, 158, 170–1, 202, 409, 413, 415–7, 519, 586, 590–1, 596  
Wronskian 98  
zero 70
- determinate  
statically 307, 333
- deviation 468, 470  
standard 468–9  
principal standard 472
- DFT *xi*, *xv*, 183, 272, 285, 289, 295
- diagonal 7, 43, 86, 104, 444, 449  
block 35, 74, 128, 171, 420, 449, 535  
main 7, 43, 52, 492  
off- 7, 32  
sub- 52, 492, 535  
super- 52, 449, 492, 535

- diagonal entry 10, 47, 70, 168, 205, 420, 445, 455–6, 600  
diagonal line 358  
diagonal matrix 7–8, 35, 41–2, 45, 85, 159, 168, 171, 204, 304, 313, 327, 400, 408, 425–6, 437, 439–40, 446, 453, 455, 472, 484, 528, 530, 608, 618, 622  
diagonal plane 362  
diagonalizable *xvi*, 403, 424, 426, 428, 437, 450, 520  
complex 427  
real 427, 432  
simultaneously 428–9  
diagonalization *xii*, 438, 456, 484  
diagonally dominant 281, 283, 421–2, 475, 498, 510, 512, 516, 584  
diamond 149  
difference  
finite 317, 521, 547  
set-theoretic *xvii*  
difference equation 341  
difference map 436  
difference polynomial 268  
differentiable 84, 136, 348  
infinitely 84  
nowhere 561  
differential delay equation 341  
differential equation *ix*, *xii–xv*, 1, 48, 106, 129, 132, 156, 183, 222, 302, 341, 351, 403, 479, 541, 556, 599  
Euler 380, 393  
homogeneous 84, 379, 381, 392, 567, 609, 621  
inhomogeneous 84, 606, 623, 627  
linear *vii*, *viii*, *xiv*, 84, 342, 376, 378  
matrix 590, 592  
ordinary *vii*, *x*, *xi*, 91, 98–9, 101, 106–7, 301, 322, 342, 376, 379–80, 385, 390, 403–4, 407, 435, 476, 479, 566–7, 570, 576, 579, 604, 606, 608, 627, 630  
partial *vii*, *ix–xii*, *xv*, *xvii*, 99, 101, 106, 129, 173, 200, 227, 230, 301, 322, 342, 376, 381, 475, 536, 542, 547, 565  
system of first order 565–7, 570–2, 577, 585, 605  
system of ordinary *ix*, *xii–xv*, 342, 530, 566, 584, 571, 579, 592, 608, 618, 630  
differential geometry 235, 381  
differential operator *xi*, 317, 341–2, 355, 376–7, 379–80, 384  
ordinary 353  
partial 381  
differentiation 348, 355–6, 594, 606  
numerical 271  
digit 297  
digital image 294, 462  
digital medium 183, 285  
digital monitor 286  
digraph 122, 311–2, 316, 327, 462  
bipartite 127  
complete 127  
connected 124  
cubical 126  
internet 126, 463, 502  
pentagonal 125  
simple 467, 502  
weighted 311, 502  
dilation coefficient 557  
dilation equation 555–6, 558, 561  
dimension *ix*, *x*, *xiii*, 76, 100, 114, 177, 341, 463  
complementary 218  
dimensional reduction 472  
dinner 626  
direct method 475, 536  
directed edge 122, 463  
directed graph — *see* digraph  
direction 630  
conjugate 543, 545, 548  
natural 631  
null 159–60, 164, 166  
principal 438–9  
disconnected 128, 463  
discontinuous 562  
discrete dynamics *xii*, 475  
discrete Fourier analysis *xi*, *xiv*, *xv*, 235, 285  
discrete Fourier representation 287, 294  
Discrete Fourier Transform *xi*, *xv*, 183, 272, 285, 289, 295  
Inverse 289  
discriminant 158, 586, 590  
discretization 317  
disk 371  
Gershgorin 420, 503  
unit 136, 371, 503  
displacement 301–2, 320, 323, 329, 608, 629  
displacement vector 302, 312, 325, 620  
distance 8, 131, 146, 235, 239, 245–6, 254, 361, 372, 467  
distinct eigenvalues 430, 453, 586  
distributed  
normally 473  
distribution 468  
distributive 8, 76, 343, 364  
divergence 86, 349  
division 48, 53, 79, 174, 261, 536  
DNR 479  
dodecahedron 127  
Dolby 293  
dollar 477, 486

domain *xvii*, 105, 342, 376, 396, 618  
 Gershgorin 420, 422  
 dominant  
   strictly diagonally 281, 283, 421–2, 475, 498, 510, 512, 516, 584  
 dominant eigenvalue 441, 495, 523–5  
 dominant eigenvector 523–4, 529  
 dominant singular value 454, 460  
 dot notation *viii*  
 dot product *xiii*, 129, 137, 146, 162, 176, 178, 193, 201, 265, 351, 365, 396, 431–3, 455, 550  
   Hermitian 178, 205, 288, 433, 444, 446  
 double eigenvalue 411, 416  
 double root 411, 576  
 doubly stochastic 505  
 downhill 236  
 drone 200  
 dual basis 350, 352, 369  
 dual linear function 369, 398  
 dual map 395  
 dual space 350, 358, 369, 395  
 dump 406  
 DVD 183  
 dyadic 561–3  
 dynamical motion *xii*, 608  
 dynamical system *viii*, *xii*, *xiii*, *xv*, 301–2, 396, 403, 407, 565, 583, 591, 603  
   infinite-dimensional 565  
 dynamics *xv*, 129, 183, 605, 628  
   continuous *xii*, 565  
   discrete *xii*, 475  
   gas 565  
   molecular 48  
   nonlinear 565, 604, 616  
   quantum 173

## E

Earth 315  
 echelon 59, 60, 62, 95, 115–6  
 economics *vii*, *ix*, 235, 407, 499  
 edge 120, 122, 311, 463, 502  
   directed 122, 463  
 eigendirection 439  
 eigenequation 408  
 eigenfunction 183  
 eigenline 429, 580, 587–8  
 eigenpolynomial 408  
 eigensolution 565–6, 572, 576, 581, 603, 610, 619  
   stable 603  
   unstable 603

eigenspace 411, 413, 424, 430, 440, 456, 493  
   generalized 631  
   zero 434  
 eigenstate 437  
 eigenvalue *vii*, *ix*, *xii–xv*, 200, 403, 408–9, 417, 420, 423, 430, 437, 440, 444–5, 447, 454, 462, 473, 475, 480, 484, 489, 522, 527, 532, 535, 539, 560, 565–6, 572, 577, 581, 586, 591, 594, 596, 608, 611  
   complete *xvi*, 412, 424, 493, 531, 577, 588  
   complex *xiv*, 412, 415–6, 421, 423, 430, 433, 496, 525, 535, 538, 578, 587  
   distinct 430, 453, 586  
   dominant 523–5  
   double 411, 416  
   generalized 443, 435, 492, 618, 630–1  
   imaginary 581, 588, 603  
   incomplete 412, 578, 581  
   intermediate 442  
   Jacobi 519  
   largest 441, 495  
   multiple 430  
   negative 582  
   positive 582  
   quadratic 493  
   real 413, 423, 430, 432, 586  
   repeated 417  
   simple 411, 416, 493, 531, 535, 537  
   smallest 441  
   subdominant 493, 502, 524, 526  
   zero 412, 421, 433–4, 581, 615  
 eigenvalue equation 408, 480, 493, 566, 610, 618  
   generalized 435, 618  
 eigenvalue matrix 530, 531  
 eigenvalue perturbation 591  
 eigenvector *ix*, *xii*, *xiii*, 403, 408–10, 429, 443, 446, 448, 473, 475, 480, 484, 503, 522, 525, 527, 537, 539, 560, 565, 609  
   co- 416, 503, 525  
   complex 413, 425, 577–8, 587, 603  
   dominant 524, 529  
   extreme 441–2  
   generalized 435, 447–8, 600, 631  
   generalized null 619  
   left 416, 503, 525  
   non-null 434, 454  
   null 433–4, 615, 618  
   orthogonal 432, 436, 611  
   orthonormal 437  
   probability 501–3  
   real 413, 490, 496, 523, 535, 537, 577–8, 588  
   unit 471, 493, 496

- eigenvector basis *xii*, *xvi*, 183, 423, 427, 432, 434, 438, 446, 448, 480, 523, 528, 566, 572, 618
- eigenvector matrix 531
- Eigenvektor 408
- Eigenwert 408
- elastic
- bar 301, 322, 608
  - beam 279
  - body 439
- elastic deformation 438
- elasticity *xii*, 358, 381
- electric charge 313
- electrical circuit *viii*, *xii*, *xiv*, 122, 129, 129, 196, 235–6, 301, 628
- electrical energy 319
- electrical engineering 173
- Electrical–Mechanical Correspondence 321, 628
- electrical network *xi*, *xv*, 120, 301, 311–2, 327, 626, 630
- electrical system 183
- electricity *xi*, 311, 315, 403
- electromagnetic wave 626
- electromagnetism 80, 173, 236, 381
- electromotive force 311
- electron 311–3
- element *xvi*
- finite 220, 235, 400, 521, 541, 547
  - real 391
  - unit 148
  - zero 76, 79, 82, 87, 140, 342
- elementary column operation 72, 74
- elementary matrix 16–7, 32, 38–9, 73, 204, 360, 362
- inverse 17, 38
- elementary reflection matrix 206, 210, 418, 532
- elementary row operation 12, 16, 23, 37, 60, 70, 418, 512
- elementary symmetric polynomial 417
- Elimination
- complex Gaussian 412
  - Gauss–Jordan 35, 41, 50
  - Gaussian *ix*–*xiv*, 1, 14, 24, 28, 40, 49, 56–7, 67, 69, 72, 102, 129, 167, 208, 237, 253, 378, 407, 409, 475, 506, 508–9, 536
  - regular Gaussian 14, 18, 171, 268
  - tridiagonal Gaussian 5, 42
- ellipse 363, 371, 438–9, 466, 487, 496, 498
- ellipsoid 363, 438–9, 465, 472
- elliptic system 542
- elongation 302–3, 309, 324, 327, 608
- elongation vector 302, 325
- ending node 311, 322
- ending vertex 122
- energy 320, 341, 583, 585
- electrical 319
- internal 309
- potential 235–6, 244, 309, 320, 583
- spectral 437
- energy function 309
- engine
- search 499, 502
- engineer *xiii*, *xviii*
- engineering *vii*, *ix*, *xiii*, 1, 156, 227, 235, 301, 313, 381, 402
- electrical 173
- entry *xvii*, 4, 592
- diagonal 10, 47, 70, 168, 205, 420, 445, 455–6, 600
  - nonzero 59, 501
  - off-diagonal 7, 32, 252, 420
  - zero 52, 59, 449, 501
- envelope 625
- equal 4
- equation 236, 506
- affine 479
  - beam 394
  - characteristic 379–80, 390, 408–9, 413, 420, 453, 475, 567, 583, 586, 621, 627
  - complex 173
  - constitutive 303
  - Daubechies 558, 561
  - difference 341
  - differential *ix*, *xii*–*xv*, 1, 48, 84, 106, 129, 132, 156, 183, 222, 302, 341, 351, 403, 479, 541, 556, 599, 625, 627
  - differential delay 341
  - dilation 555–6, 558, 561
  - eigenvalue 408, 480, 493, 566, 610, 618
  - equilibrium 314
  - Euler 380, 393
  - Fibonacci 481, 486–7
  - fixed-point 506, 509, 546, 559, 563
  - Fredholm integral 377
  - functional 556
  - generalized characteristic 435, 618
  - Haar 555, 563
  - heat 394
  - homogeneous differential 84, 379, 381, 392, 567, 609, 621
  - inhomogeneous differential 84, 606, 623, 627
  - inhomogeneous iterative 479
  - integral *vii*, 76, 106, 183, 341–2, 376–8, 556
  - integro-differential 629
  - iterative 476–8, 479
  - linear differential *vii*, *viii*, *xiv*, 84, 342, 376, 378

- equation (*continued*)  
 Laplace 381, 383, 385, 393  
 matrix differential 590, 592  
 Newtonian 618, 621  
 normal 247, 251, 272, 458  
 ordinary differential *vii*, *x*, *xi*, 91, 98–9,  
 101, 106–7, 301, 322, 342, 376, 379–80,  
 385, 390, 403–4, 407, 435, 476, 479,  
 566–7, 570, 576, 579, 604, 606, 608,  
 627, 630  
 partial differential *vii*, *ix–xii*, *xv*, *xvii*, 99,  
 101, 106, 129, 173, 200, 227, 230, 301,  
 322, 342, 376, 381, 475, 536, 542, 547,  
 565  
 Poisson 385, 390, 521  
 polynomial 416  
 quadratic 64, 166, 621  
 Schrödinger 173, 394  
 system of first order differential 565–7,  
 570–2, 577, 585, 605  
 system of ordinary differential *ix*, *xii–xv*,  
 342, 530, 566, 584, 571, 579, 592, 608,  
 618, 630  
 Volterra integral 378  
 weighted normal 247, 252, 256, 317  
 equilateral 328, 500, 619  
 equilibrium *ix*, *xi*, 236, 301, 309, 403, 476,  
 565, 581, 587, 605, 618, 621, 629  
 stable 235–6, 301–2, 579, 590, 605, 615  
 unstable 235–6, 301, 590  
 equilibrium equation 314  
 equilibrium mechanics 235  
 equilibrium point 568, 579  
 equilibrium solution 301, 405, 476, 479, 488,  
 493, 565, 579, 597, 622  
 equivalence class 87  
 equivalence relation 87  
 equivalent 2, 87, 150  
 equivalent norm 150, 152  
 error *ix*, 237, 254  
 experimental 237, 467  
 least squares 235, 251–2, 271, 458  
 maximal 261  
 measurement 256, 470  
 numerical 249, 523, 536  
 round-off *x*, 55, 199, 206, 544  
 squared 255–6, 260, 272, 274  
 weighted 252, 256  
 error correcting code 464  
 error-free *xix*  
 error function 274  
 error vector 254, 508, 514  
 Euclidean geometry 99, 130, 137, 203  
 Euclidean isometry 373  
 Euclidean matrix norm 460–1, 497  
 Euclidean norm *xiii*, 130, 142, 172, 174, 224,  
 236, 250, 455, 458, 460, 468, 473, 489,  
 524, 532, 538, 544, 546  
 Euclidean space *x*, *xi*, 75–7, 94, 99, 130, 146,  
 341, 403, 426, 600  
 Euler angle 203  
 Euler differential equation 380, 393  
 Euler–Rodrigues formula 602  
 Euler’s formula 125, 175, 392  
 evaluation 341  
 even 86–7, 286  
 executive 504  
 exercise *xvi*  
 existence 1, 380, 383–4, 401, 479, 566, 593, 610  
 expander graph 464  
 expected value 468  
 expenditure 258  
 experiment 254, 293  
 experimental data 237–8  
 experimental error 237, 467  
 exponential  
   complex 175, 180, 183, 192, 285, 287, 390,  
   549  
   matrix *ix*, *xii*, 565, 593–4, 597, 599, 601–2,  
   606  
   sampled 285–7, 415, 436  
 exponential ansatz 379, 390, 567, 576, 621  
 exponential decay 405, 565, 580–1, 621  
 exponential function 175, 261, 264, 275, 277,  
 403–4, 428, 565–6, 578, 622  
 exponential growth 258–9, 565, 586, 603–4  
 exponential series 596, 599, 601–2  
 exponential solution 381, 408  
 external force 110, 301, 309, 320, 335, 384,  
 388, 565, 605, 608, 623, 627, 630  
 extinction 406  
 extreme eigenvector 441–2
- F**
- face 126, 467  
 factor 415  
   shear 361  
 factorization  
   Cholesky 171  
   Gaussian 1, 419, 437  
    $LDL^T$  *xi*, 45, 167, 437, 542  
    $LDV$  41  
    $LU$  *x*, *xiv*, *xvi*, 1, 18, 20, 41, 50, 70, 268,  
   501, 536, 542  
   matrix 1, 171, 183, 205, 536  
    $MM^T$  171  
   permuted  $LDV$  42  
   permuted  $LU$  27–8, 60, 70  
    $QR$  *xi*, 205, 210, 522, 529, 539  
   spectral 437

- farmer 504  
 Fast Fourier Transform *xi*, 235, 296  
 father 504  
 fault-tolerant 464  
 FBI 555  
 FFT *xi*, 235, 296  
 Fibonacci equation 481, 486–7  
 Fibonacci integer 482–3, 485  
 Fibonacci matrix 412, 428  
 Fibonacci number 481, 483, 486  
 field 76
  - skew 364
  - vector 81, 574
 finance *vii*, *ix*, 1, 403, 407, 475  
 fingerprint 555  
 finite difference 317, 521, 547  
 finite-dimensional *xiii*, *xiv*, 101, 149, 213, 219, 248, 356  
 finite element 220, 235, 400, 521, 541, 547  
 first order system 605  
 fitting
  - curve 283
  - data *xiii*, 132, 237, 254
 fixed point 493, 506, 509, 546, 559, 563
  - stable 493
 floating point 48, 58  
 floor 322  
 flow 313
  - fluid 574
  - gradient 581, 585
 flower 482, 504  
 fluctuation 467, 470  
 fluid flow 574  
 fluid mechanics 48, 80, 173, 236, 381, 400, 565  
 focus 587–9, 591  
 FOM *xii*, 476, 541, 546, 570  
 font *viii*, 203, 283  
 force *viii*, 110, 565
  - electromotive 311
  - external 110, 301, 309, 320, 335, 384, 388, 565, 605, 608, 623, 627, 630
  - frictional 621–2
  - gravitational 80, 259, 311
  - internal 303, 320, 327, 333
  - mechanical 327
  - periodic *xii*, 565, 623–4, 626, 630–1
  - vibrational 630
 force balance 304, 327, 333  
 force vector 327  
 forcing frequency 624, 630  
 fork
  - tuning 624
 form
  - bilinear 156
  - canonical 368
  - Jordan canonical *xii*, *xiii*, 403, 447, 450, 490, 525, 598
  - linear 161
  - Minkowski 375
  - phase-amplitude 89, 610
  - quadratic *xi*, 86, 157–61, 166–7, 170, 241, 245, 346, 437, 440–2, 583
  - row echelon 59, 60, 62, 95, 115–6
  - triangular 2, 14
 formal adjoint 396  
 formula *xvii*
  - addition 89
  - Binet 483, 485
  - binomial 176, 393
  - Euler 125, 175, 392
  - Euler–Rodrigues 602
  - Gram–Schmidt 194
  - orthogonal basis 189, 611
  - orthonormal basis 188
  - polarization 160
  - Pythagorean 188, 460
  - Rodrigues 228, 277
  - quadratic 166
  - Sherman–Morrison–Woodbury 35
 formulation
  - Galerkin *xii*, 200
 FORTRAN 14  
 Forward Substitution *xiii*, *xiv*, 3, 20, 49, 53, 282, 518  
 foundation *xiii*  
 Fourier analysis *ix*, *x*, *xii*, *xv*, 75, 78, 99, 135, 173, 180, 183, 188, 227, 285, 287, 476, 555
  - discrete *xi*, *xiv*, *xv*, 235, 285
 Fourier basis function 549  
 Fourier coefficient 289, 291, 294, 296–7, 470  
 Fourier series 91, 191, 549, 553  
 Fourier transform 376, 559
  - discrete *xi*, *xv*, 183, 272, 285, 289, 295
  - fast *xi*, 235, 296
 fragment
  - skeletal 393
 framework 120, 322  
 Fredholm Alternative *vii*, *ix*, *xi*, 183, 222, 226, 312, 330, 352, 377, 631  
 Fredholm integral equation 377  
 free space 394  
 free variable 62, 63, 67, 96, 108, 119–20, 315  
 frequency 273, 578, 609, 623, 629
  - audible 287, 614
  - forcing 624, 630
  - high- 183, 287, 291, 294, 555

- frequency (*continued*)
   
low- 291, 294
   
natural 565, 611, 624–5, 630–1
   
resonant 626–7, 632
   
vibrational 610, 621, 624
- friction *xii*, 302, 565, 608, 620, 626–7, 629
- friction matrix 622
- frictional coefficient 499, 504
- frictional force 621–2
- Frobenius norm 156, 466
- fruit 482
- full pivoting 57
- Full Orthogonalization Method *xii*, 476, 541, 546
- function *viii*, *xiv*, *xvii*, *xviii*, 75, 78–9, 285, 341, 343, 440, 475
   
adjoint 396–8
   
affine *xi*, 245, 239, 343, 370, 555
   
algebraic 166
   
analytic 84, 87, 91
   
basis 549, 562
   
bilinear 347, 354
   
box 549, 555, 559
   
coefficient 353
   
complex exponential 175, 180, 183, 192, 285, 287, 390, 549
   
complex linear 342
   
complex trigonometric 176–7
   
complex-valued 129, 173, 177, 179, 391
   
constant 78
   
continuous *xi*, 83, 133, 150, 179, 219–20, 274, 347, 349, 377, 559
   
continuously differentiable 84, 136, 379
   
cosine *xvii*, 138, 175–6, 183, 285, 581, 609–10
   
Daubechies 559–60
   
discontinuous 562
   
dual 369, 398
   
energy 309
   
error 274
   
even 86–7
   
exponential 175, 261, 264, 275, 277, 403–4, 428, 565–6, 578, 622
   
Fourier basis 549
   
generalized 351
   
Haar 549, 555, 559
   
Hamiltonian 583
   
harmonic 381
   
hat 556, 563
   
homogeneous 161
   
hyperbolic 176
   
identity 343, 355
   
infinitely differentiable 84
   
integrable 84
   
inverse linear 355
   
invertible linear 387
- function (*continued*)
   
linear *ix*, *xi*–*xiv*, *xvi*, *xvii*, 239, 341–2, 349–50, 352, 355, 358, 369–70, 378, 383, 395–6, 599
   
matrix-valued 592, 594, 606
   
mean zero 84
   
non-analytic 84, 87
   
nonlinear 324, 341
   
nowhere differentiable 561
   
odd 87
   
orthogonal *xi*, 183, 559–60
   
periodic 86, 611
   
piecewise constant 551
   
piecewise cubic 263, 269, 279
   
positive definite 398–9, 401
   
power 320
   
quadratic *xi*, *xiii*, 235, 239–41, 259, 274, 401, 545, 582–3
   
quasi-periodic 565, 611
   
rational 166, 261, 442
   
real linear 342, 391
   
sample 79–80, 235, 285
   
scaling 399, 549, 555, 558, 559–60, 563
   
self-adjoint 398–9, 436
   
sinc 273
   
sine *xvii*, 176, 183, 269, 285, 581, 610
   
skew-adjoint 400
   
smooth 84
   
special 200
   
translation 346
   
trigonometric *xi*, *xiv*, *xvii*, 89, 164, 175–6, 183, 235, 272, 292, 578, 580–1
   
unit 148
   
vector-valued 80, 98, 136, 341, 605
   
wave 173, 341
   
zero 79, 83, 134, 343, 405
- function evaluation 341
- function space *x*, *xiii*, *xv*, 79, 80, 83, 133, 146, 163, 185, 190, 220, 224, 301, 341, 396, 401, 541
- function theory 135
- functional analysis *xii*
- functional equation 556
- fundamental subspace 114, 183, 221
- Fundamental Theorem of Algebra 98, 124, 415
- Fundamental Theorem of Calculus 347, 356, 606
- Fundamental Theorem of Linear Algebra 114, 461
- Fundamental Theorem of Principal Component Analysis 472

## G

- Galerkin formulation *xii*, 200

- game *xii*, 200, 341, 375  
 gas dynamics 565  
 Gauss–Jordan Elimination 35, 41, 50  
 Gauss–Seidel matrix 514–5, 519  
 Gauss–Seidel Method *xii*, 475, 512, 514, 517, 519–20  
 Gauss–Seidel spectral radius 520  
 Gaussian Elimination *ix–xiv*, 1, 14, 24, 28, 40, 49, 56–7, 67, 69, 72, 102, 129, 167, 208, 237, 253, 378, 407, 409, 475, 506, 508–9, 536  
 complex 412  
 regular 14, 18, 171, 268  
 tridiagonal 5, 42  
 Gaussian factorization 1, 419, 437  
 general position 375  
 general relativity *vii*, 341  
 general solution 91, 107, 111, 480, 606, 618, 625  
 generalized characteristic equation 435, 618  
 generalized eigenspace 631  
 generalized eigenvalue 443, 435, 492, 618, 630–1  
 generalized eigenvector 435, 447–8, 600, 619, 631  
 generalized function 351  
 Generalized Minimal Residual Method *xii*, 476, 546–7, 549  
 generator 629  
 infinitesimal 599–602  
 generic 48  
 genetics 475, 504  
 genotype 504  
 geodesic 235  
 geodetic survey 171  
 geometric mean 148  
 geometric modeling 279  
 geometric multiplicity 424  
 geometric series 499  
 geometry *viii*, *xi*, *xii*, 10, 75, 99, 129, 183, 200, 202, 238, 341, 358, 464, 472, 565, 599  
 differential 235, 381  
 Euclidean 99, 130, 137, 203  
 geophysical image 102  
 Gershgorin Circle Theorem 420, 475, 503  
 Gershgorin disk 420, 503  
 Gershgorin domain 420, 422  
 Gibbs phenomenon 562  
 glide reflection 375  
 global minimum 240  
 globally asymptotically stable 405, 488–90, 492–3, 579, 622  
 globally stable 405, 579  
 GMRES *xii*, 476, 546–7, 549  
 golden ratio 483  
 Google 126, 463, 502  
 gradient 349, 545, 582  
 conjugate 476, 542, 544–5, 548  
 gradient descent 545, 549  
 gradient flow 581, 585  
 Gram matrix 129, 161–3, 182, 246, 255, 274, 301, 309, 316, 327, 351, 398, 403, 439, 454, 456, 462, 470, 543, 622, 630  
 weighted 163, 247  
 Gram–Schmidt formula 194  
 Gram–Schmidt process *ix*, *xi*, *xv*, 183, 192, 194–5, 198–9, 205, 208, 215, 227, 231, 249, 266, 445, 475, 527, 529, 538  
 stable 199, 538  
 graph 120, 303, 311, 317, 463  
 complete 127, 464  
 connected 121, 144  
 directed — *see* digraph  
 disconnected 128, 463  
 expander 464  
 planar 125  
 random 463  
 simple 120, 311, 463  
 spectral *xiv*, *xv*, 462–3  
 weighted 311  
 graph Laplacian *xv*, 301, 317–8, 462, 464  
 graph spectrum 462, 467  
 graph theory *viii*, *x*, *xiv*, 126  
 graphics *xi*  
 computer *viii*, *xi*, 52, 200, 203, 279, 283, 286, 341, 358, 374–5, 565, 599  
 gravitational constant 259  
 gravitational force 80, 259, 311  
 gravitational potential 311  
 gravity 235, 259, 302, 307, 327, 333, 583, 613  
 gray scale 470  
 grid 317, 521–2  
 ground 327  
 grounded 315, 318  
 group 202, 204, 599, 600  
 one-parameter 599–603  
 orthogonal 203  
 group theory *xii*, 464, 599  
 growth  
 exponential 258–9, 565, 586, 603–4  
 one-parameter 599–603  
 polynomial 580, 604  
 spiral 483  
 guess 475  
 initial 544, 548  
 inspired 379, 567

## H

- $H^1$  inner product 136, 144, 233  
 $H^1$  norm 136

- Haar equation 555, 563  
 Haar function 549, 555, 559  
 Haar wavelet 549–50, 552–3, 555, 562  
 half-life 257, 404, 406  
 half-open 79  
 half-plane 580–1  
 half-sampled 297  
 Hamiltonian function 583  
 Hamiltonian system 583, 585  
 hard spring 303  
 hardware 57  
 harmonic function 381  
 harmonic polynomial 381, 393  
 hat function 556, 563  
 health informatics 467  
 heat 319, 394, 626  
 heat equation 394  
 height 259  
 Heisenberg commutation relation 355  
 helix 85, 374  
 Hermite polynomial 233  
 Hermitian adjoint 181, 205  
 Hermitian dot product 178, 205, 288, 433, 444, 446  
 Hermitian inner product 180–1, 192  
     weighted 435  
 Hermitian matrix 181–2, 435, 439  
 Hermitian norm 445, 489  
 Hermitian transpose 205, 444  
 hertz 614  
 Hessenberg matrix 535–6, 539, 542  
 Hessian matrix 242  
 hexagon 292, 619–20  
 hi fi 287  
 high-dimensional data 471  
 high-frequency 183, 287, 291, 294, 555  
 high precision 48, 57  
 high-resolution image 462  
 higher order system 605  
 Hilbert matrix 57–8, 164, 212, 276–7, 465, 516, 548, 584  
 Hilbert space 135, 341  
 hill 236, 583  
 hiss 293  
 hole 125  
 home 258  
 homogeneous 15, 67, 144, 379, 405  
 homogeneous differential equation 84, 379, 381, 392, 567, 609, 621  
 homogeneous function 161  
 homogeneous solution 388  
 homogeneous system *vii*, *xi*, *xii*, 67, 95, 99, 106, 108, 342, 376, 378, 384, 388, 394, 409, 571, 585  
 Hooke's Law 303, 306, 309, 322, 327, 625  
 house 258, 632  
 Householder matrix 206, 210–1, 532, 535  
 Householder Method 209, 211–2, 532  
 Householder reflection 535, 539  
 Householder vector 210–1, 536  
 hunter 406, 479  
 hydrogen 620  
 hyperbola 569, 570  
 hyperbolic function 176  
 hyperbolic rotation 375  
 hyperplane 103, 362
- I**
- I-beam 322  
 icosahedron 127  
 idempotent 16, 109, 216, 419  
 identity  
     additive 76  
     Jacobi 10, 354, 602  
     trigonometric 175  
 identity function 343, 355  
 identity matrix 7, 8, 16, 25, 31, 70, 200, 409, 588, 593, 618  
 identity permutation 26, 415  
 identity transformation 348, 429  
 IDFT 289  
 ill-conditioned 56–7, 211, 249, 276–7, 461–2, 525  
 image *x*, *xvi*, 75, 105, 107–8, 114, 116–7, 221, 223–4, 237, 294, 312, 383, 429, 434, 457  
     digital 294, 462  
     geophysical 102  
     high-resolution 462  
     JPEG 555  
     medical 102, 295  
     still 102, 285  
 image processing *vii*, *viii*, *x–xii*, *xv*, 1, 48, 99, 102, 183, 188, 404, 467, 470, 555  
 image vector 473  
 imaginary axis 580–1  
 imaginary eigenvalue 581, 588, 603  
 imaginary part 173, 177, 287, 391, 394, 575–6  
 imaginary unit *xvii*, 173  
 implicit iterative system 492  
 improper isometry 373  
 improper node 588–9, 591, 604  
 improper orthogonal matrix 202, 358, 438  
 inbreeding 504  
 incidence matrix 122, 124, 128, 303, 312, 314, 317, 325, 327, 462, 616, 622  
     reduced 303–4, 316, 318, 330, 335, 339, 622  
 incompatible *xi*, 62  
 incomplete eigenvalue 412, 578, 581

- incomplete matrix 403, 424, 480, 490, 575, 594, 603  
 inconsistent 62  
 increase  
   steepest 545, 582  
 indefinite 159, 583  
 indefinite integral *xviii*, 348, 356  
 independent  
   linearly *ix*, *x*, *xiii*, 75, 93, 95–6, 99, 100, 161, 185, 341, 380, 423, 448, 570, 599  
   statistically 470  
 indeterminate  
   statically 306, 315, 334  
 index 357, 447  
 inductance 626, 629  
 induction *x*, 434, 451  
 inductor 311, 628  
 inequality *ix*, 129  
   Cauchy-Schwarz 129, 137, 142–3, 179, 469  
   complex 177  
   integral 143  
   Minkowski 145–6  
   product 154  
   Sylvester 120  
   triangle 129, 142–4, 146, 154, 179, 498  
 inertia 439  
 infantry 625  
 infinite 79  
 infinite-dimensional *x*, *xiv*, *xv*, 101, 129, 133, 149, 151, 213, 219–20, 274, 301, 341, 349, 351, 355–6, 396, 401, 541  
 infinite-dimensional dynamical system 565  
 infinite-dimensional subspace 219  
 $\infty$  matrix norm 495–6, 499, 510, 515  
 $\infty$  norm 145, 151, 245, 255, 473, 489, 496, 514, 524, 544  
 infinite series 394  
 infinitely differentiable 84  
 infinitesimal generator 599–602  
 infinitesimal motion 616  
 inflection point 240  
 informatics 467  
 inhomogeneity 110  
 inhomogeneous differential equation 84, 606, 623, 627  
 inhomogeneous iterative equation 479  
 inhomogeneous system *vii*, *xi*, *xii*, 67, 106, 110–1, 342, 376, 383–4, 388, 394, 565, 585, 605–6, 630  
 initial condition 404–5, 480, 570–2, 593, 609–10  
 initial guess 544, 548  
 initial position 609–10, 612  
 initial time 621  
 initial value problem 376, 386, 570, 594, 598, 606  
 initial vector 475, 540  
 initial velocity 609–10, 618, 622  
 inner product *ix–xi*, *xiii*, *xiv*, 129–30, 133, 137, 144, 156–7, 163, 179, 237, 245, 347, 350, 395  
   complex 184  
 $H^1$  136, 144, 233  
 Hermitian 180–1, 192, 435  
 $L^2$  133, 135, 180, 182, 185, 191, 219, 227, 232, 234, 274, 550–1, 557, 560  
 real 184, 395, 401  
 Sobolev 136, 144, 233  
 weighted 131, 135, 182, 246, 265, 309, 396, 435, 543, 618  
 inner product matrix 156  
 inner product space 130, 140, 161, 183–4, 193, 196, 213, 219, 245, 342, 347, 350, 358  
   dual 358  
 inspired guess 379, 567  
 instructor *xviii*  
 instrument 467, 626  
 integer *xvii*, 561  
   Fibonacci 481–3, 485–6  
   random 487  
   tribonacci 486  
 integrable 84  
 integral *x*, 557  
   definite *xviii*  
   line 125  
   indefinite *xviii*, 348, 356  
   Lebesgue 135  
   trigonometric 175, 177, 624  
   weighted 182  
 integral equation *vii*, 76, 106, 183, 341–2, 376–8, 556  
   Fredholm 377  
   Volterra 378  
 integral inequality 143  
 integral operator *xi*, 341  
 integral transform 376  
 integration 347, 606  
   numerical 271, 562  
 integration constant 404  
 integration operator 347  
 integro-differential equation 629  
 interchange 209  
   column 57  
   row 23, 25, 56, 70, 361  
 interchange matrix 25  
 interconnectedness 120  
 interest 476, 478–9  
 intermediate eigenvalue 442  
 internal energy 309  
 internal force 303, 320, 327, 333

- internal motion 388, 627  
 internal power 319  
 internal vibration 565, 624  
 internet 126, 258, 463, 475, 499, 502  
 interpolant 263  
 interpolation *viii, ix, xi, xiii, 52, 79, 227, 235, 262*  
     polynomial 235, 260, 262, 271, 279  
     sinc 273  
     spline 52, 235  
     trigonometric 86, 235, 287, 293  
 intersection *xvii, 65*  
 interval *xvii, 79, 83–4, 219, 386, 557*  
     closed *xvii, 146*  
     open 146  
     time 476  
 invariant subspace *xv, 429–31, 452, 487, 492, 548, 603–4*  
     complex 430, 452  
     real 452  
 inverse *xiii, xiv, 1, 17, 31, 32, 50, 111, 200, 355*  
     additive 8, 76  
     left 31, 36, 38, 356  
     pseudo- 403, 457, 467  
     right 31, 35, 356  
 Inverse Discrete Fourier Transform 289  
 inverse elementary matrix 17, 38  
 inverse linear function 355  
 inverse matrix *x, 17, 31–3, 38, 40, 44, 72, 102, 111, 428, 457*  
 inverse permutation 32, 34  
 Inverse Power Method 526  
     Shifted 526–7, 534, 539  
 inversion 79, 102  
 invertible 33, 106, 387, 421, 439  
 IQ 467  
 irrational number 611  
 irregular 505  
 irrotational 86  
 island 504  
 isometry 372–3  
     improper 373  
     proper 373, 419  
 isomorphic 356  
 isotope 257, 404  
 iterate 476, 506  
 iteration *xv, 475*  
     linear *vii, 403, 475*  
     nonlinear 53, 475  
 iterative equation 476–8, 479  
 iterative method *xv, 475, 506, 536*  
     linear *vii, 403*  
     naive 517  
     semi-direct *xii, 475, 536, 547*  
 iterative system 53, 563  
     affine 488  
     implicit 492  
     nonlinear 53, 475  
     order of 481, 493  
     second order 493
- J**
- Jacobi eigenvalue 519  
 Jacobi identity 10, 354, 602  
 Jacobi matrix 509, 511, 515, 519  
 Jacobi Method *xii, 475, 509–11, 513, 517, 519–20*  
 Jacobi spectral radius 520  
 Jacobian matrix 605  
 jar 623  
 jagged 286  
 JAVA 14  
 joint 120, 322–3  
 Jordan basis 448, 450–1, 453, 480, 488, 576–7  
 Jordan Basis Theorem 448, 450  
 Jordan block 416–7, 449–50, 453, 598  
 Jordan canonical form *xii, xiii, 403, 447, 450, 490, 525, 598*  
 Jordan chain 447–8, 450–2, 488, 576, 579, 603–4  
     null 447, 451  
 Jordan chain solution 576–7, 581  
 JPEG 555  
 junction 311
- K**
- kernel *x, 75, 105, 107–8, 114–5, 117, 124–5, 221, 223–4, 331, 378, 380, 384, 411, 429, 434, 456, 463, 631*  
 kill 479  
 Kirchhoff's Current Law 313–4, 317  
 Kirchhoff's Voltage Law 312, 314  
 Krylov approximation 541–2  
 Krylov subspace *xv, 475, 536–7, 539–40, 546, 549*  
 Krylov vector 537, 547  
 Ky Fan norm 466
- L**
- $L^1$  norm 145, 147, 153, 182, 274  
 $L^2$  Hermitian inner product 180  
 $L^2$  inner product 133, 135, 182, 185, 191, 219, 227, 232, 234, 274, 550–1, 557, 560  
 $L^2$  norm 133, 145, 152–3, 185, 191  
 $L^2$  squared error 274  
 $L^p$  norm 145  
 $L^\infty$  norm 145, 147, 152–3, 182

- laborer 504  
 Lagrange data 284  
 Lagrange multiplier 441  
 Lagrange polynomial 262, 284  
 Lagrangian notation *viii*  
 Laguerre polynomial 231, 234, 279  
 Lanczos Method *xii*, 475, 539  
 language 404  
 Laplace equation 381, 383, 385, 393  
 Laplace transform 376  
 Laplacian 349, 354, 381, 393  
     graph *xv*, 301, 317–8, 462, 464  
 large 188, 475  
 largest eigenvalue 441, 495, 523–5  
 laser printing 283  
 lattice 505  
 Law 139  
     Hooke's 303, 306, 309, 322, 327, 625  
     Kirchhoff's Current 313–4, 317  
     Kirchhoff's Voltage 312, 314  
     Newton's 565, 608  
     Ohm's 312, 319  
     Voltage Balance 312, 314, 629  
 Law of Cosines 139  
 LC circuit 630  
 $LDL^T$  factorization *xi*, 45, 167, 437, 542  
 $LDV$  factorization 41  
     permuted 42  
 leading coefficient 367  
 leaf 482  
 learning *vii*, 235, 404, 467  
 least squares *ix*, *xiii*–*xv*, 129, 132, 230, 235, 255, 266, 468  
     weighted 252, 256, 265  
 least squares approximation 188, 263, 272  
 least squares coefficient 266  
 least squares error 235, 251–2, 271, 458  
     weighted 252, 256  
 least squares line 474  
 least squares minimizer 183, 237  
 least squares solution *ix*, *xi*, 237–8, 250–1, 317, 403, 458  
 Lebesgue integral 135  
 left eigenvector 416, 503, 525  
 left half-plane 580–1  
 left-handed basis 103, 202, 222  
 left inverse 31, 36, 38, 356  
 left limit *xviii*  
 left null space *x*, 113  
 Legendre polynomial 232, 234, 277–8  
 Leibniz rule 594  
 Leibnizian notation *viii*  
 length 120, 130, 323  
 letter 283  
 level curve 585  
 level set 585  
 license 479  
 light  
     speed of 159  
     stroboscopic 286  
 light cone 159–60  
 light ray 160, 235  
 limit *xviii*  
 line 65, 83, 87–8, 237, 239, 254, 259–60, 301, 343, 363, 370–1, 587, 615  
 least squares 474  
 parallel 371  
 spectral energy 437  
 stable 587, 589, 591  
 tangent 341, 600  
 unstable 587, 589  
 line integral 125  
 line segment 473  
 linear *vii*, *ix*, 2, 342  
 linear algebra *vii*, *xi*, *xiii*, *xv*, 1, 75, 114, 126, 183, 243, 341, 403, 506  
     Fundamental Theorem of 114, 461  
     numerical 48  
 linear algebraic system *vii*, *ix*, 341–2, 376, 386, 506, 517, 540  
 linear analysis *ix*, *x*  
 linear approximation 324, 329, 341, 388  
 linear combination 87, 95, 101, 287, 342, 388, 599, 618  
 linear control system *vii*, *xv*, 376  
 linear differential equation *vii*, *viii*, *xiv*, 84, 342, 376, 378  
 linear differential operator *xi*, 317, 341–2, 355, 376–7, 379–80, 384  
 linear dynamical system *xii*, 565, 603  
 linear form 161  
 linear function *ix*, *xi*–*xiv*, *xvi*, *xvii*, 239, 341–2, 349–50, 352, 355, 358, 369–70, 378, 383, 395–6, 599  
     adjoint 396–7  
     dual 369, 398  
     inverse 355  
     invertible 387  
     positive definite 398–9, 401  
     real 342, 391  
     self-adjoint 398–9, 436  
     skew-adjoint 400  
 linear independence *ix*, *x*, *xiii*, 75, 99, 177, 341, 570  
 linear integral equation *vii*, 76, 106, 183, 341–2, 376–7, 556  
 linear integral operator *xi*, 341  
 linear iteration *vii*, 403, 475i  
 linear iterative system *vii*, *ix*, *xii*–*xv*, 476, 479, 493, 499, 500, 522, 560, 584, 605  
 linear map 341–2

- linear mathematics *ix*  
 linear motion 588–90, 616, 618  
 linear operator 75, 156, 341–3, 347, 376, 437, 541  
 linear polynomial 187  
 linear programming 235  
 linear superposition *vii*, *ix*, *xi*, *xv*, 110, 222, 235, 250, 262, 342, 378, 480, 565, 630  
 linear system *vii*, *ix*, *xi*, 4, 6, 20, 23, 40, 59, 63, 67, 75, 99, 105–7, 376, 461, 475, 541, 565, 571, 577  
 adjoint 112  
 compatible *ix*, *xi*, 8, 11, 62, 224  
 complex *xiv*, 566  
 equivalent 2  
 forced 565  
 homogeneous *vii*, *xi*, *xii*, 67, 95, 99, 106, 108, 342, 376, 378, 384, 388, 394, 409, 571, 585  
 ill-conditioned 57, 211, 461  
 incompatible *xi*, 62  
 inconsistent 62  
 inhomogeneous *vii*, *xi*, *xii*, 106, 110–1, 342, 376, 383–4, 388, 394, 565, 585, 605–6, 630  
 lower triangular 3, 20  
 singular 461  
 sparse *xv*, 52, 475, 536  
 triangular 2, 29, 197, 542  
 weak 132, 398, 540  
 linear system of ordinary differential equations *ix*, *xii*–*xv*, 342, 530, 566, 584, 571, 608, 630  
 second order *xii*, 618  
 linear system operation 2, 23, 37  
 linear transformation *ix*, *xiii*, *xiv*, 341–2, 358, 403, 426, 429, 457, 554, 599  
 self-adjoint 436  
 linearity *ix*, 2  
 linearization 341, 605  
 linearly dependent 93, 95–6, 100, 571  
 linearly independent *ix*, *x*, *xiii*, 75, 93, 95–6, 99, 100, 161, 185, 341, 380, 423, 448, 570, 594, 599  
 local minimum 236, 242, 441  
 localized 549  
 logarithm *xvii*, 258, 269, 599  
 London 626  
 loop 120  
 low-frequency 291, 294  
 lower bidiagonal 52  
 lower triangular *xvi*, 3, 16–7, 20, 39, 73, 518  
 special *xvi*  
 strictly *xvi*, 16–8, 28, 39, 41–2, 45, 60, 85, 168, 509, 530  
 lower unitriangular *xvi*, 16–8, 20, 28, 39, 41–3, 45, 60, 85, 168, 530  
 LP 287  
 LU factorization *x*, *xiv*, *xvi*, 1, 18, 20, 41, 50, 70, 268, 501, 536, 542  
 permuted 27–8, 60, 70  
 Lucas number 486
- M**
- machine learning *vii*, 235, 404, 467  
 magic square 104  
 magnitude 630  
 main diagonal 7, 43, 52, 492  
 manifold 235, 605  
 manufacturing 235  
 map 342  
 difference 436  
 linear 341–2  
 perspective 374–5  
 scaling 399  
 shift 415, 436  
 zero 361  
 MAPLE 14, 57  
 market 475  
 Markov chain *xii*, 475, 499–502  
 Markov process *ix*, *xiv*, 463–4, 563  
 mass *viii*, 110, 236, 301, 311, 341, 609, 615, 621–3, 629  
 center of 439  
 mass matrix 396, 608, 616, 618, 620, 622, 630  
 mass–spring chain *xi*, *xiv*, 301, 309, 317, 399, 403, 565, 608, 610, 619, 628, 630  
 mass–spring ring 339  
 MATHEMATICA 14, 57, 409  
 mathematician *xviii*  
 mathematics 1, 75, 227, 314, 381  
 applied *vii*, *ix*, *x*, 1, 48, 230, 475  
 financial 1  
 MATLAB 14, 409  
 matrix *ix*–*xi*, *xiii*, *xiv*, *xvii*, 1, 3, 48, 75, 105, 133, 223, 341, 343, 407, 445, 457, 475  
 adjacency 317  
 adjoint 396  
 affine 372, 603  
 approximating 462  
 Arnoldi 540  
 augmented 12, 24, 36, 60, 66–7  
 banded 55  
 bidiagonal 52, 536  
 block 11, 35, 603  
 block diagonal 35, 74, 128, 171, 420, 449, 535, 598  
 block upper triangular 74, 535  
 circulant 282, 436

- matrix (*continued*)  
 coefficient 4, 6, 63, 157, 224, 235, 241, 343, 476, 479, 484, 499, 508, 528, 531, 566, 575, 591, 606, 608, 618, 630  
 cofactor 112  
 commuting 10, 601  
 complete *xvi*, 403, 424–6, 428, 430–2, 444, 450, 480, 484, 490, 493, 522, 566, 572, 575, 603  
 complex 5, 181, 212, 226, 536, 566  
 complex diagonalizable 427  
 conductance 313, 317  
 convergent 488–9, 495–6, 508, 517  
 covariance 163, 470–1, 473  
 data 462, 467, 470–1, 473  
 deflated 420  
 degree 317  
 diagonal 7–8, 35, 41–2, 45, 85, 159, 168, 171, 204, 304, 313, 327, 400, 408, 425–6, 437, 439–40, 446, 453, 455, 472, 484, 528, 530, 608, 618, 622  
 diagonalizable *xvi*, 403, 424, 426, 428, 437, 450, 520  
 doubly stochastic 505  
 eigenvalue 530, 531  
 eigenvector 531  
 elementary 16–7, 32, 38–9, 73, 204, 360, 362  
 elementary reflection 206, 210, 418, 532  
 Fibonacci 412, 428  
 friction 622  
 Gauss–Seidel 514–5, 519  
 generic 48  
 Gram 129, 161–3, 182, 246, 255, 274, 301, 309, 316, 327, 351, 398, 403, 439, 454, 456, 462, 470, 543, 622, 630  
 graph Laplacian *xv*, 301, 317–8, 462, 464  
 Hermitian 181–2, 435, 439  
 Hessenberg 535–6, 539, 542  
 Hessian 242  
 Hilbert 57–8, 164, 212, 276–7, 465, 516, 548, 584  
 Householder 206, 210–1, 532, 535  
 idempotent 16, 109, 216, 419  
 identity 7, 8, 16, 25, 31, 70, 200, 409, 588, 593, 618  
 ill-conditioned 56–7, 276–7, 461, 525  
 improper orthogonal 202, 358, 438  
 incidence 122, 124, 128, 303, 312, 314, 317, 325, 327, 462, 616, 622  
 incomplete 403, 424, 480, 490, 575, 594, 603  
 indefinite 159  
 inner product 156  
 interchange 25  
 matrix (*continued*)  
 inverse *x*, 17, 31–3, 38, 40, 44, 72, 102, 111, 428, 457  
 invertible 33, 106, 421, 439  
 irregular 505  
 Jacobi 509, 511, 515, 519  
 Jacobian 605  
 Jordan block 416–7, 449–50, 453, 598  
 linearization 605  
 lower bidiagonal 52  
 lower triangular *xvi*, 16–7, 20, 39, 73, 518  
 lower unitriangular *xvi*, 16–8, 20, 28, 39, 41–3, 45, 60, 85, 168, 530  
 mass 396, 608, 616, 618, 620, 622, 630  
 negative definite 159–60, 171, 581, 583  
 negative semi-definite 159  
 nilpotent 16, 418, 453  
 nonsingular *xi*, 23–4, 28, 32, 39, 42, 44, 62, 85, 99, 106, 204, 367, 380, 422, 457, 460, 492, 599, 630  
 non-square 60, 403  
 non-symmetric 157  
 normal 44, 446  
 normalized data 470, 473  
 orthogonal *xiii*, 183, 200, 202, 205, 208, 210, 358, 373, 413, 431, 437, 439, 444, 446, 457, 530, 552  
 orthogonal projection 216, 440  
 orthonormal 201  
 pentadiagonal 516  
 perfect *xvi*, 424  
 permutation 25, 27–8, 32, 42, 45, 60, 71–2, 74, 97, 204–5, 419, 430  
 positive definite *vii*, *ix*, *xi–xiv*, 129, 156–7, 159–61, 164–5, 167, 170–1, 181–2, 204, 235, 241–2, 244, 246, 252, 301, 304, 309, 313, 316, 327, 396, 398, 432, 439, 443, 473, 528, 531, 542, 544, 581, 583, 608, 618, 622  
 positive semi-definite *xi*, 158, 161, 182, 244–5, 301, 316, 320, 433, 454, 470, 473, 514, 615, 618, 622  
 positive upper triangular 205, 529–30  
 projection 216, 440  
 proper orthogonal 202–3, 205, 222, 358, 438–9, 600  
 pseudoinverse 403, 457, 467  
 quadratic coefficient 241  
 rank one 66  
 real 5, 77, 425, 430, 440, 444, 446, 476, 536, 575–6, 595  
 real diagonalizable 427, 432  
 rectangular 31, 453, 457  
 reduced incidence 303–4, 316, 318, 330, 335, 339, 622  
 reduced resistivity 316

- matrix (*continued*)
   
 reflection 206, 210, 418, 532
   
 regular *xvi*, 13, 18, 42, 45, 52, 70, 85, 501, 530–1, 536, 542
   
 regular transition 501
   
 resistance 313
   
 resistivity 314, 316, 320
   
 rotation 34, 358, 414, 430
   
 row echelon 59, 60, 62, 95, 115–6
   
 self-adjoint 399, 618
   
 semi-simple *xvi*, 424
   
 shift 436
   
 similar 73, 367, 418, 425–6, 428, 465, 498, 532, 575, 598
   
 simultaneously diagonalizable 428–9
   
 singular 23, 70, 314, 403, 409, 411–2, 597, 631
   
 skew-symmetric 47, 73, 85–6, 204, 400, 435, 439, 600–2
   
 SOR 475, 518–9
   
 sparse 48, 536, 548
   
 special lower triangular *xvi*
  
 special orthogonal 222
   
 special upper triangular *xvi*
  
 square 4, 18, 23, 31, 33, 45, 403, 416, 426, 453, 457, 495, 542, 596
   
 stiffness 305, 309, 320, 327, 611, 615, 618, 622
   
 strictly diagonally dominant 281, 283, 421–2, 475, 498, 510, 512, 516, 584
   
 strictly lower triangular *xvi*, 16–8, 28, 39, 41–2, 45, 60, 85, 168, 509, 530
   
 strictly upper triangular 16, 85, 509
   
 symmetric *xi*, *xiv*, 45, 85–6, 167, 171, 183, 208, 216, 226, 398–9, 403, 432, 434, 437, 440–1, 446, 454, 465, 487, 532, 537, 542, 581, 585, 631
   
 symmetric tridiagonal 532
   
 transition 499–501, 505, 525, 528, 536, 598
   
 transposed 72, 112, 162, 304, 395
   
 tricirculant 54, 282–3, 420, 436
   
 tridiagonal 52, 281, 304, 419, 492, 512, 526, 532, 535–6, 539, 542
   
 unipotent 16–7
   
 unitary 205, 212, 439, 444–6
   
 unitriangular *xvi*, 16–8, 20, 28, 38–9, 41–3, 45, 60, 85, 168, 530, 543
   
 upper bidiagonal 52
   
 upper Hessenberg 535–6, 539, 542
   
 upper triangular *xvi*, 13, 16, 23–4, 28, 39, 70, 204–5, 210, 425, 428, 444–6, 465, 518, 527, 530, 532, 602
   
 upper unitriangular *xvi*, 16, 18, 38, 41–3, 543
   
 Vandermonde 20, 74, 260, 268
   
 wavelet 204, 224, 552, 554
- matrix (*continued*)
   
 weight 256
   
 weighted Gram 163, 247
   
 Young 519–20, 579
   
 zero 7, 8, 61, 77, 361, 457, 488, 597
   
 matrix addition 5, 8, 12, 43, 349
   
 matrix algebra 7, 99
   
 matrix arithmetic *xiii*, 1, 8, 77
   
 matrix differential equation 590, 592
   
 matrix exponential *ix*, *xiii*, 565, 593–4, 597, 599, 601–2, 606
   
 matrix factorization 1, 18, 20, 41, 45, 50, 171, 183, 205, 536
   
 matrix logarithm 599
   
 matrix multiplication 5, 8, 43, 48, 51, 95, 106, 223–4, 343, 352, 355, 365, 397, 403, 429, 457, 475
   
 matrix norm *xi*, *xii*, *xv*, 153–4, 156, 460–1, 466, 475–6, 495–7, 496–9, 510, 515, 596
   
 Euclidean 460–1, 497
   
 Frobenius 156, 466
   
  $\infty$  495–6, 499, 510, 515
   
 Ky Fan 466
   
 natural 153–4, 495, 499
   
 matrix pair 618, 630
   
 matrix polynomial 11, 453
   
 matrix power 475, 479, 484, 488, 502
   
 matrix product 33, 72, 130
   
 matrix pseudoinverse 403, 457, 467
   
 matrix series 499, 596
   
 matrix solution 572
   
 matrix square root 439, 465, 620
   
 matrix-valued function 592, 594, 606
   
 max norm 145, 151
   
 maximal error 261
   
 maximal rank 456
   
 maximization principle 235, 443
   
 constrained 442–3
   
 maximum *xvii*, 150, 235, 240, 441, 442
   
 mean 10, 84, 148, 348, 467, 470, 473
   
 arithmetic 148
   
 geometric 148
   
 mean zero 10, 84, 468, 470
   
 measure theory 135
   
 measurement 254, 256, 467–70
   
 measurement error 256, 470
   
 mechanical force 327
   
 mechanical structure 301
   
 mechanical vibration 565
   
 mechanics *viii*, *xi*, *xii*, 129, 156, 183, 236, 342, 396, 439, 565, 627
   
 classical 341, 388, 583
   
 continuum *xi*, 235–6, 351, 399
   
 equilibrium 235

- mechanics (*continued*)  
 fluid 48, 80, 173, 236, 381, 400, 565  
 quantum *vii*, *viii*, *x*, 10, 48, 129, 135, 173,  
 183, 200, 202, 227, 341, 349, 355, 381,  
 388, 437, 467, 583, 599  
 relativistic 341, 388  
 rigid body 200, 203  
 solid 236  
 mechanism 301, 331, 336, 599, 616, 618  
 medical image 102, 295  
 medium *ix*, 183, 285  
 memory 56, 513, 561  
 mesh point 279  
 methane 139  
 method  
 Arnoldi *xii*, 475, 538  
 Back Substitution *x*, *xiii*, *xiv*, 3, 14, 21, 24,  
 41, 50, 53, 62, 208, 211, 282, 518  
 Conjugate Gradient 476, 542, 544–5, 548  
 deflation 420, 526  
 direct 475, 536  
 Forward Substitution *xiii*, *xiv*, 3, 20, 49,  
 53, 282, 518  
 Full Orthogonalization (FOM) *xii*, 476,  
 541, 546  
 Gauss–Seidel *xii*, 475, 512, 514, 517, 519–20  
 Gaussian Elimination *ix*–*xiv*, 1, 14, 24, 28,  
 40, 49, 56–7, 67, 69, 72, 102, 129, 167,  
 208, 237, 253, 378, 407, 409, 412, 475,  
 506, 508–9, 536  
 Generalized Minimal Residual (GMRES)  
*xii*, 476, 546–7, 549  
 Gram–Schmidt *ix*, *xi*, *xv*, 183, 192, 194–5,  
 198–9, 205, 208, 215, 227, 231, 249,  
 266, 445, 475, 527, 529, 538  
 Householder 209, 211–2, 532  
 Inverse Power 526  
 iterative *vii*, *xv*, 403, 475, 506, 536  
 Jacobi *xii*, 475, 509–11, 513, 517, 519–20  
 Lanczos *xii*, 475, 539  
 nave iterative 517  
 Power *xii*, 475, 522, 524, 529, 536–7, 568  
 regular Gaussian Elimination 14, 18, 171,  
 268  
 semi-direct *xii*, 475, 536, 547  
 Shifted Inverse Power 526–7, 534, 539  
 Singular Value Decomposition *xii*, 403,  
 455, 457, 461, 473  
 Strassen 51  
 Successive Over–Relaxation *xii*, 475, 517–20  
 undetermined coefficient 372, 385–6, 500,  
 623  
 tridiagonal elimination 5, 42  
 metric 159  
 microwave 626  
 Midpoint Rule 271  
 milk 486  
 minimal polynomial 453, 537  
 minimization *xiii*, *xv*, 129, 156, 238, 241, 309,  
 546, 583  
 minimization principle *vii*, *ix*, 235–6, 320,  
 342, 402  
 minimization problem *xi*–*xiv*, 255  
 minimizer 183, 237, 241, 401, 545  
 minimum *xvii*, 150, 235–6, 240  
 global 240  
 local 236, 242, 441  
 minimum norm solution 224, 458  
 mining  
 data 467  
 Minkowski form 375  
 Minkowski inequality 145–6  
 Minkowski metric 159  
 Minkowski space-time 160  
 Minneapolis 501  
 Minnesota 406, 479, 501, 546  
 missile 269  
 missing data 471  
 $MM^T$  factorization 171  
 mode  
 normal *xii*, 565, 611  
 unstable 615, 618–9  
 vibrational 616  
 model 620  
 modeling *viii*, 235, 279, 309  
 modular arithmetic *xvii*  
 modulate 624  
 modulus 174, 177, 489  
 molasses 622  
 molecule 139, 437, 608  
 benzene 620  
 carbon tetrachloride 620  
 triatomic 616, 619, 632  
 water 620, 626, 632  
 molecular dynamics 48  
 moment 330, 439  
 momentum 341, 355  
 money 476, 486  
 monic polynomial 227, 453  
 monitor 286  
 monomial 89, 94, 98, 100, 163–4, 186, 231,  
 265, 271, 275  
 complex 393  
 sampled 265  
 trigonometric 190  
 monomial polynomial 268  
 monomial sample vector 265  
 month 477  
 mother wavelet 550, 552, 555–6, 558

- motion 388, 403, 608, 631  
 circular 616  
 damped 621  
 dynamical *xii*, 608  
 infinitesimal 616  
 internal 388, 627  
 linear 588–90, 616, 618  
 nonlinear 600  
 periodic 621, 624  
 rigid 301, 327, 335, 373, 599, 601, 616  
 screw *xi*, 341, 373, 419, 602  
 movie *xii*, 200, 285, 287, 293, 341  
 MP3 183  
 multiple eigenvalue 430  
 multiplication 5, 48–9, 53, 79, 261, 348, 457, 536  
 complex 173, 296, 298  
 matrix 5, 8, 43, 48, 51, 95, 106, 223–4, 343, 352, 355, 365, 397, 403, 429, 457, 475  
 noncommutative 6, 26, 355, 360, 364, 601  
 quaternion 364  
 real 296  
 scalar 5, 8, 43, 76, 78, 87, 343, 349, 390  
 multiplicative property 596  
 multiplicative unit 7  
 multiplicity 416–7, 424, 454, 581  
 algebraic 424  
 geometric 424  
 multiplier  
 Lagrange 441  
 multipole 547  
 multivariable calculus *x*, 235, 242–3, 342, 441, 545, 582  
 music 287, 626
- N**
- naïve iterative method 517  
 natural boundary condition 280, 283–4  
 natural direction 631  
 natural frequency 565, 611, 624–5, 630–1  
 natural matrix norm 153–4, 495, 499  
 natural vibration 614  
 Nature 235, 317, 320, 483  
 negative definite 159–60, 171, 581, 583  
 negative eigenvalue 582  
 negative semi-definite 159  
 network *viii*, *xv*, 301, 312, 315, 317–8, 320, 463, 499  
 communication 464  
 electrical *xi*, *xv*, 120, 301, 311–2, 327, 626, 630  
 newton 112  
 Newton difference polynomial 268  
 Newtonian equation 618, 621  
 Newtonian notation *viii*  
 Newtonian physics *vii*  
 Newtonian system 614  
 Newton's Law 565, 608  
*n*-gon — *see* polygon  
 nilpotent 16, 418, 453  
 node 120, 311, 313, 315, 317, 339  
 ending 311  
 improper 588–9, 591, 604  
 stable 586, 588–9, 591  
 starting 311  
 terminating 311, 322  
 unstable 587–9, 591  
 noise 183, 293, 555  
 non-analytic function 84, 87  
 non-autonomous 570, 598  
 noncommutative 6, 26, 355, 360, 364, 601  
 non-coplanar 88  
 nonlinear *vii*, 255, 341, 388, 611, 616  
 nonlinear dynamics 565, 604, 616  
 nonlinear function 324, 341  
 nonlinear iteration 53, 475  
 nonlinear motion 600  
 nonlinear system 64, 66, 342, 475, 568, 604  
 nonnegative orthant 83  
 non-null eigenvector 434, 454  
 non-Pythagorean 131  
 non-resonant 628  
 nonsingular *xi*, 23–4, 28, 32, 39, 42, 44, 62, 85, 99, 106, 204, 367, 380, 422, 457, 460, 492, 599, 630  
 non-square matrix 60, 403  
 non-symmetric matrix 157  
 nontrivial solution 67, 95  
 nonzero vector 95  
 norm *ix–xi*, *xiii*, *xiv*, 129, 131, 135, 137, 142, 144, 146, 174, 188–9, 237, 245, 489, 495, 581  
 convergence in 151  
 equivalent 150, 152  
 Euclidean *xiii*, 130, 142, 172, 174, 224, 236, 250, 455, 458, 460, 468, 473, 489, 524, 532, 538, 544, 546  
 Euclidean matrix 460–1, 497  
 Frobenius 156, 466  
 $H^1$  136  
 Hermitian 445, 489  
 $\infty$  matrix 495–6, 499, 510, 515  
 $\infty$  145, 151, 245, 255, 473, 489, 496, 514, 524, 544  
 Ky Fan 466  
 $L^1$  145, 147, 153, 182, 274  
 $L^2$  133, 145, 152–3, 185, 191  
 $L^\infty$  145, 147, 152–3, 182  
 matrix *xi*, *xii*, *xv*, 153–4, 156, 460–1, 466, 475–6, 495–7, 496–9, 510, 515, 596

- norm (*continued*)  
 max 145, 151  
 minimum 224, 458  
 natural matrix 153–4, 495, 499  
 1– 145, 245, 255, 466  
 residual 237  
 Sobolev 136  
 2– 145  
 weighted 131, 135, 237, 252, 468
- normal equation 247, 251, 272, 458  
 weighted 247, 252, 256, 317
- normal matrix 44, 446
- normal mode *xii*, 565, 611
- normal vector 217
- normalize 468, 470, 473
- normally distributed 473
- normed vector space 144, 372
- north pole 474
- notation  
 dot *viii*  
 Lagrangian *viii*  
 Leibnizian *viii*  
 Newtonian *viii*  
 prime *viii*
- nowhere differentiable 561
- nuclear reactor 406
- nucleus 437
- null direction 159–60, 164, 166
- null eigenvector 433–4, 615, 618–9
- null space *x*, 106  
 left *x*, 113
- number *viii*, 3, 5, 78  
 complex *xvii*, 80, 129, 173  
 condition 57, 460, 466  
 dyadic 561, 563  
 Fibonacci 481–3, 485–6  
 irrational 611  
 Lucas 486  
 pseudo-random 464, 487  
 random 295, 464  
 rational *xvii*, 611  
 real *xvii*, 78, 137, 173, 563  
 spectral condition 525, 591  
 tribonacci 486
- numerical algorithm *viii–xi*, 48, 129, 183, 199, 400, 403, 475, 536, 547
- numerical analysis *vii*, *ix*, *xii*, *xiii*, *xv*, 1, 75, 78, 132, 156, 227, 230, 233, 235, 271, 279, 317, 475
- numerical approximation 220, 235, 403, 416, 467
- numerical artifact 206
- numerical differentiation 271
- numerical error 249, 523, 536
- numerical integration 271, 562
- numerical linear algebra 48
- O**
- object 259
- observable 341
- occupation 504
- octahedron 127, 149
- odd 87, 286
- off-diagonal 7, 32, 252, 420
- offspring 479, 482
- Ohm's Law 312, 319
- oil 623, 628
- 1 norm 145, 245, 255, 466
- one-parameter group 599–603
- open 79, 136, 146, 151  
 half- 79
- Open Rule 271
- operation  
 arithmetic 48, 199, 212, 534, 536, 548  
 elementary column 72, 74  
 elementary row 12, 16, 23, 37, 60, 70, 418, 512  
 linear system 2, 23, 37
- operator  
 derivative 348, 353  
 differential *xi*, 317, 341–2, 355, 376–7, 379–80, 384  
 differentiation 348, 355  
 integral *xi*, 341  
 integration 347  
 Laplacian 349, 354, 381, 393  
 linear 75, 156, 341–3, 347, 376, 437, 541  
 ordinary differential 353  
 partial differential 381  
 quantum mechanical *xi*  
 self-adjoint 183  
 Schrödinger 437
- optics 235, 375
- optimization 235, 441–2, 466  
 constrained 441
- orange juice 486
- orbit 568
- order 379, 481, 493, 567  
 first 565–7, 570–2, 577, 585, 605  
 higher 605  
 reduction of 379, 390  
 second *xii*, 618  
 stabilization 537, 540, 547, 549
- ordinary derivative *xviii*
- ordinary differential equation *vii*, *x*, *xi*, 91, 98–9, 101, 106–7, 301, 322, 342, 376, 379–80, 385, 390, 403–4, 407, 435, 476, 479, 566–7, 570, 576, 579, 604, 606, 608, 627, 630
- homogeneous 390

ordinary differential equation (*continued*)  
 inhomogeneous 606  
 system of *ix*, *xii–xv*, 342, 530, 566, 584, 571, 579, 592, 608, 630  
 ordinary differential operator 353  
 orientation 122, 201, 311, 313, 317  
 origin 88, 343  
 ornamentation 322  
 orthant 83  
 orthogonal *x*, 140, 184–5, 189, 213, 216, 222, 335, 540, 549, 618, 631  
 orthogonal basis *xi*, *xiii*, *xv*, 184, 189, 194, 201, 214, 235, 266, 403, 435, 446, 551, 611, 618  
 orthogonal basis formula 189, 611  
 orthogonal complement 217–9, 221, 431, 631  
 orthogonal eigenvector 432, 436, 611  
 orthogonal function *xi*, 183, 559–60  
 orthogonal group 203  
 orthogonal matrix *xiii*, 183, 200, 202, 205, 208, 210, 358, 373, 413, 431, 437, 439, 444, 446, 457, 530, 552  
 improper 202, 358, 438  
 special 222  
 proper 202–3, 205, 222, 358, 438–9, 600  
 orthogonal polynomial *xi*, *xiv*, 141, 183, 186, 227–8, 276–7  
 orthogonal projection *xi*, *xiii*, *xv*, 183, 213, 216, 218, 223, 235, 248, 361–2, 440, 457, 471–2, 539, 631  
 orthogonal subspace *xv*, 183, 216  
 orthogonal system 552  
 orthogonal vector *xiii–xv*, 140, 185  
 orthogonality *vii*, *xi*, 184, 235, 287, 295, 312, 476, 558, 562  
 orthogonalization 475  
 orthonormal basis 184, 188, 194–6, 198–9, 201, 204, 213, 235, 248, 288, 432, 437–8, 444, 456–7, 460, 475, 528–9, 538  
 orthonormal basis formula 188  
 orthonormal column 444, 455–6  
 orthonormal eigenvector 437  
 orthonormal matrix 201  
 orthonormal rows 455  
 orthonormalize 529, 539  
 orthonormality *ix*, 184  
 oscillation 626  
 out of plane 627  
 outer space 301, 327  
 oven 258, 626  
 overdamped 622, 629  
 overflow 524  
 over-relaxation *xii*, 475, 517–20  
 oxygen 620

## P

*p* norm 145, 245  
 Padé approximation 261  
 page 126, 463, 502  
 PageRank 463, 502  
 pair  
 matrix 618, 630  
 parabola 15, 240, 259–60, 590–1  
 paraboloid 83  
 parachutist 269  
 parallel 65, 83, 88, 93, 137, 142, 147–8, 187, 371, 500  
 parallel computer 513  
 parallelizable 513  
 parallelogram 140, 344, 361  
 parameter 599  
 Cayley–Klein 203  
 relaxation 518  
 variation of 385, 606, 623  
 part  
 imaginary 173, 177, 287, 391, 394, 575–6  
 real 173, 177, 287, 365, 391, 394, 575–6, 581, 591, 603  
 partial derivative *xviii*, 242, 349  
 partial differential equation *vii*, *ix–xii*, *xv*, *xvii*, 99, 101, 106, 129, 173, 200, 227, 230, 301, 322, 342, 376, 381, 475, 536, 542, 547, 565  
 partial differential operator 381  
 partial pivoting 56, 62  
 partial sum 554  
 particular solution 107, 384, 606, 623–5, 630  
 partitioning 463  
 path 121  
 PCA *ix*, *xii*, *xiv*, *xv*, 255, 403, 467, 471  
 peak 90  
 peg 279  
 pendulum 236  
 pentadiagonal matrix 516  
 pentagon 125, 288, 467  
 perfect matrix *xvi*, 424  
 perfect square 166  
 period 611, 621  
 period 2 solution 495  
 periodic 172, 285, 565, 587  
 periodic boundary condition 280, 283–4  
 periodic force *xii*, 565, 623–4, 626, 630–1  
 periodic function 86, 611  
 periodic motion 621, 624  
 periodic spline 280, 283  
 periodic vibration 618  
 permutation 26–7, 45, 56, 72, 428  
 column 418  
 identity 26, 415  
 inverse 32, 34

- permutation (*continued*)
   
  row 428
   
  sign of 72
- permutation matrix 25, 27–8, 32, 42, 45, 60, 71–2, 74, 97, 204–5, 419, 430
- permuted *LDV* factorization 42
- permuted *LU* factorization 27–8, 60, 70
- perpendicular 140, 202, 255
- Perron–Frobenius Theorem 501
- perspective map 374–5
- perturbation 523, 525, 591
- phase *xviii*, 174, 627
- phase-amplitude 89, 587, 610
- phase lag 627
- phase plane 567–8, 576, 586, 605
- phase portrait *xii*, 568, 586–90, 623
- phase shift 89, 273, 587, 609
- phenomenon
   
  Gibbs 562
- physical model 620
- physics *vii*, *ix*, *x*, *xii*, 1, 200, 202, 227, 235, 301, 314, 327, 341–2, 381, 402, 407, 437, 499, 565
- continuum 156
- Newtonian *vii*
- statistical 464
- piecewise constant 551
- piecewise cubic 263, 269, 279–80
- pivot 12, 14, 18, 22–3, 28, 41, 49, 56, 59, 61, 70, 114, 167, 412
- pivot column 56
- pivoting 23, 55, 57, 62
   
  full 57
   
  partial 56, 62
- pixel 470
- planar graph 125
- planar system 565, 585
- planar vector field 81, 86
- plane 64, 82, 88, 250, 259, 358
   
  complex 173, 420, 580
   
  coordinate 362
   
  diagonal 362
   
  left half- 580–1
   
  out of 627
   
  phase 567–8, 576, 586, 605
- plane curve 86
- planet 259, 341
- plant 504
- platform 616
- Platonic solid 127
- plot
   
  scatter 469, 471, 478
- point 65, 83, 87–8, 235, 250, 570, 587
   
  boundary 503
   
  closest *xi*, 183, 235, 238, 245–6, 298
- point (*continued*)
   
  critical 240, 242
   
  data *xi*, 235, 237, 254–5, 272, 283, 467, 470, 474
   
  dyadic 561–2
   
  equilibrium 568, 579
   
  fixed 493, 506, 509, 546, 559, 563
   
  floating 48, 58
   
  inflection 240
   
  mesh 279
   
  saddle 587, 589
   
  sample 79, 105, 285
   
  singular *xv*, 380
- pointer 56
- Poisson equation 385, 390, 521
- polar angle 174
- polar coordinate 90, 136, 174
- polar decomposition 439
- polarization 160
- pole 151, 474
- police 196, 254
- polygon 125, 208, 288, 467
- polynomial *xi*, *xiv*, 75, 78, 83, 89, 91, 94, 98, 100, 114, 139, 219, 260–1, 413, 440, 578, 581, 603
- approximating 266, 279
- characteristic 408, 415, 453, 475
- Chebyshev 233
- complete monomial 268
- constant 78
- cubic 260, 267, 280
- elementary symmetric 417
- even 86
- factored 415
- harmonic 381, 393
- Hermite 233
- interpolating 260, 262, 271, 279
- Lagrange 262, 284
- Laguerre 231, 234, 279
- Legendre 232, 234, 277–8
- linear 187
- matrix 11, 453
- minimal 453, 537
- monic 227, 453
- Newton difference 268
- orthogonal *xi*, *xiv*, 141, 183, 186, 227–8, 276–7
- piecewise cubic 263, 269, 279–80
- quadratic *xvi*, 167, 185, 190, 235, 240, 260, 267
- quartic 221, 264, 269, 276
- quintic 233
- radial 277
- sampled 265
- symmetric 417

- polynomial (*continued*)  
 Taylor 269, 324, 383  
 trigonometric 75, 90–1, 94, 176, 190, 273  
 unit 148  
 zero 78
- polynomial algebra 78
- polynomial equation 416
- polynomial growth 580, 604
- polynomial interpolation 235, 260, 262, 271, 279
- polytope 149
- population 259, 475, 479, 482, 487, 504
- portrait  
 phase *xii*, 568, 586–90, 623
- position 254, 355  
 general 375  
 initial 609–10, 612
- positive definite *vii*, *ix*, *xi–xiv*, 129, 156–7, 159–61, 164–5, 167, 170–1, 181–2, 204, 235, 241–2, 244, 246, 252, 301, 304, 309, 313, 316, 327, 396, 398, 432, 439, 443, 473, 528, 531, 542, 544, 581, 583, 608, 618, 622
- positive semi-definite *xi*, 158, 161, 182, 244–5, 301, 316, 320, 433, 454, 470, 473, 514, 615, 618, 622
- positive upper triangular 205, 529–30
- positivity 130, 133, 144, 146, 156
- potential  
 gravitational 311  
 voltage 311–2, 315, 318, 320
- potential energy 235–6, 244, 309, 320, 583
- potential theory 173
- power 235, 319–20  
 matrix 475, 479, 484, 488, 502
- power ansatz 380, 479
- power function 320
- Power Method *xii*, 475, 522, 524, 529, 536–7, 568
- Inverse 526  
  Shifted Inverse 526–7, 534, 539
- power series 175
- precision 48, 57, 461
- predator 479
- prestressed 320
- price 258
- prime notation *viii*
- primitive root of unity 288
- principal axis 465, 472, 487
- Principal Component Analysis *ix*, *xii*, *xiv*, *xv*, 255, 403, 467, 471–2  
  Fundamental Theorem of 472
- principal coordinate 472, 477
- principal direction 471–4
- principal standard deviation 472
- principal stretch 438–9
- principal variance 472–3
- principle  
  maximization 235, 442–3  
  minimization *vii*, *ix*, 235–6, 320, 342, 402  
  optimization 235, 441–2, 466  
  Reality 391, 484  
  superposition 75, 106, 111, 378, 388  
  Uncertainty 355
- printing 283
- probabilistic process 479
- probability *vii*, *viii*, *xii–xiv*, 463, 475, 499  
  transitional 501–2
- probability distribution 468
- probability eigenvector 501–3
- probability vector 473, 500–1
- problem *viii*  
  boundary value *vii*, *x*, *xi*, *xv*, 54, 75, 92, 99, 136, 183, 222, 235, 322, 342, 376–6, 386, 389, 397, 399, 541–2  
  initial value 376, 386, 570, 594, 598, 606  
  minimization *xi–xiv*, 255
- process *viii*, *xii*, *xiv*, 403, 475, 499  
  Gram–Schmidt *ix*, *xi*, *xv*, 183, 192, 194–5, 198, 205, 208, 215, 227, 231, 249, 266, 445, 475, 527, 529, 538
- Markov *ix*, *xiv*, 463–4, 563
- probabilistic 479
- stable Gram–Schmidt 199, 538
- stochastic 499
- processing  
  image *vii*, *viii*, *x–xii*, *xv*, 1, 48, 99, 102, 183, 188, 404, 467, 470, 555  
  signal *vii*, *viii*, *x–xiii*, 1, 75, 80, 99, 102, 129, 183, 188, 235, 272, 293, 476  
  video 48, 102, 188, 200, 285, 294–5, 341
- processor 56, 513
- product *xvii*, 256  
  Cartesian 81, 86, 133, 347, 377  
  complex inner 184  
  cross 140, 187, 239, 305, 602  
  dot *xiii*, 129, 137, 146, 162, 176, 178, 193, 201, 265, 351, 365, 396, 431–3, 455, 550  
   $H^1$  inner 136, 144, 233  
  Hermitian dot 178, 205, 288, 433, 444, 446  
  Hermitian inner 180–1, 192, 435  
  inner *ix–xi*, *xiii*, *xiv*, 129–30, 133, 137, 144, 156–7, 163, 179, 237, 245, 347, 350, 395  
   $L^2$  inner 133, 135, 180, 182, 185, 191, 219, 227, 232, 234, 274, 550–1, 557, 560  
  matrix 33, 72, 130  
  real inner 184, 395, 401  
  Sobolev inner 136, 144, 233  
  vector 6, 130

- product (*continued*)  
 weighted inner 131, 135, 182, 246, 265,  
 309, 396, 435, 543, 618
- product inequality 154
- professional 504
- profit 258
- programming  
 computer 14, 28  
 linear 235
- projection 341, 353, 374, 467  
 orthogonal *xi*, *xiii*, *xv*, 183, 213, 216, 218,  
 223, 235, 248, 361–2, 440, 457, 471–2,  
 539, 631  
 random 471
- projection matrix 216, 440
- proof *viii*
- proper isometry 373, 419
- proper orthogonal matrix 202–3, 205, 222,  
 358, 438–9, 600
- proper subspace 210
- proper value 408
- proper vector 408
- property  
 multiplicative 596
- pseudo-random number 464, 487
- pseudocode 14, 24, 31, 49, 56, 206, 212, 536,  
 544
- pseudoinverse 403, 457, 467
- Pythagorean formula 188, 460
- Pythagorean Theorem 130–2
- Q**
- Q.E.D. *xvii*
- QR* algorithm *xii*, 200, 475, 527, 529, 531–2,  
 535–6, 538
- QR* factorization *xi*, 205, 210, 522, 529, 539
- quadrant 81
- quadratic coefficient matrix 241
- quadratic eigenvalue 493
- quadratic equation 64, 166, 621
- quadratic form *xi*, 86, 157, 161, 166–7, 170,  
 241, 245, 346, 437, 440–2, 583  
 indefinite 159  
 negative definite 159–60  
 negative semi-definite 159  
 positive definite 157, 160  
 positive semi-definite 158
- quadratic formula 166
- quadratic function *xi*, *xiii*, 235, 239–41, 259,  
 274, 401, 545, 582–3
- quadratic minimization problem *xi*–xiv
- quadratic polynomial *xvi*, 167, 185, 190, 235,  
 240, 260, 267
- quantize 475, 583
- quantum dynamics 173
- quantum mechanics *vii*, *viii*, *x*, *xi*, 10, 48,  
 129, 135, 173, 183, 200, 202, 227, 341,  
 349, 355, 381, 388, 437, 467, 583, 599
- quantum mechanical operator *xi*
- quartic polynomial 221, 264, 269, 276
- quasi-periodic *ix*, *xii*, 565, 611, 617–9, 624,  
 630–1
- quaternion 364–5, 583
- quaternion conjugate 365
- quintic polynomial 233
- quotient  
 Rayleigh 442
- quotient space 87, 105, 357
- R**
- rabbit 482, 487
- radial coordinate 383
- radial polynomial 277
- radian *xvii*
- radical 416
- radio 322, 293
- radioactive 257, 404, 406
- radium 259
- radius  
 spectral *xii*, 475, 490, 492–3, 495–8, 508,  
 518, 520, 524  
 unit *viii*
- random graph 463
- random integer 487
- random number 295, 464
- random projection 471
- random walk 463
- range *xvi*, 105, 342
- range finder 269
- rank 61–3, 66, 96, 99, 108, 114, 124, 162,  
 165, 223, 365, 434, 455, 457, 461, 465  
 maximal 456
- rank  $k$  approximation 462
- rank one matrix 66
- rate  
 decay 257, 404, 622  
 sample 287
- ratio  
 golden 483
- rational arithmetic 58
- rational function 166, 261, 442
- rational number *xvii*, 611
- ray 160, 235
- Rayleigh quotient 442
- reactor 406
- real 177, 390
- real addition 296
- real analysis 151
- real arithmetic 173
- real basis 575

- real diagonalizable 427, 432  
 real eigenvalue 413, 423, 430, 432, 586  
 real eigenvector 413, 490, 496, 523, 535, 537, 577–8, 588  
 real element 391  
 real inner product 184, 395, 401  
 real linear function 342, 391  
 real matrix 5, 77, 425, 430, 440, 444, 446, 476, 536, 575–6, 595  
 real multiplication 296  
 real number *xvii*, 78, 137, 173, 563  
 real part 173, 177, 287, 365, 391, 394, 575–6, 581, 591, 603  
 real scalar 5, 177, 476  
 real solution *xiv*, 577  
 real subspace 452  
 real vector 76, 391  
 real vector space *xi*, 76, 342  
 Reality Principle 391, 484  
 recipe *viii*  
 reciprocal 31  
 reciprocity relation 322  
 recognition 285, 404, 467, 555  
 reconstruction 299, 555  
 record 287, 293  
 rectangle 361  
 rectangular 3, 31, 453, 457  
 rectangular grid 81  
 reduced incidence matrix 303–4, 316, 318, 330, 335, 339, 622  
 reduction of order 379, 390  
 refined Gershgorin domain 422  
 reflection *xi*, 206, 341, 353, 358, 360–2, 373, 399, 440, 457  
     elementary 206, 210, 418, 532  
     glide 375  
     Householder 535, 539  
 reflection matrix 206, 210, 418, 532  
 regiment 625  
 region  
     stability 590  
 regular *xvi*  
 regular Gaussian Elimination 14, 18, 171, 268  
 regular matrix *xvi*, 13, 18, 42, 45, 52, 70, 85, 501, 530–1, 536, 542  
 regular transition matrix 501  
 reinforced 334, 336  
 relation  
     constitutive 302, 312, 327  
     equivalence 87  
     Heisenberg commutation 355  
     reciprocity 322  
     relativity *vii*, 34, 235–6, 341, 388  
         general *vii*, 34  
         special 159, 358, 375  
 relaxation  
     over- *xii*, 475, 517–20  
     under- 518  
 relaxation parameter 518  
 repeated eigenvalue 417  
 repeated root 576, 622  
 rescale 562  
 resident 504  
 residual norm 237  
 residual vector 237, 522, 541, 544–5, 548  
 resistance 259, 311, 313, 319, 628–9  
 resistance matrix 313  
 resistanceless 608  
 resistivity matrix 314, 316, 320  
 resistor 301, 313, 628  
 resonance *viii*, *ix*, *xii*, 565, 625–6, 630–1  
 resonant ansatz 631  
 resonant frequency 626–7, 632  
 resonant solution 628, 631  
 resonant vibration 565, 625  
 response 384, 606, 624  
 retina 375  
 retrieval 48  
 reversal  
     bit 297  
     time 569  
 RGB 470  
 right angle 140, 184  
 right-hand side 4, 12, 15, 49  
 right-handed basis 103, 201–2, 222  
 right inverse 31, 35, 356  
 right limit *xviii*  
 rigid body mechanics 200, 203  
 rigid motion 301, 327, 335, 373, 599, 601, 616  
 ring 339  
 rivet 323  
 RLC circuit 626, 629  
 robot 324, 616  
 rod 322  
 Rodrigues formula 228, 277  
 Rolle's Theorem 231  
 roller 339  
 roof 322  
 root 98, 231, 379–80, 413, 415, 567  
     complex 114, 390, 621  
     double 411, 576  
     matrix square 439, 465, 620  
     repeated 576, 622  
     simple 411  
 square *xvii*, 12, 131, 166, 171, 175, 185, 198, 214, 269, 454, 468, 611

- root of unity 288  
  primitive 288, 292, 296
- rotation *xi*, 200–1, 286, 313, 329, 331, 341, 344, 353, 373, 399, 419, 429, 438–9, 457, 600–1, 616, 618  
  clockwise 360  
  counterclockwise 359, 600  
  hyperbolic 375
- rotation matrix 34, 358, 414, 430
- round-off error *x*, 55, 199, 206, 544
- routing 463, 499
- row 4, 6, 7, 12, 43, 70, 114, 470, 523–4, 536  
  orthonormal 455  
  zero 70, 73
- row echelon 59, 60, 62, 95, 115–6
- row interchange 23, 25, 56, 70, 361
- row operation 12, 16, 23, 37, 60, 70, 418, 512
- row permutation 428
- row pointer 56
- row space *x*, 113
- row sum 10, 419, 502  
  absolute 155, 496, 498, 510
- row vector 4, 6, 130, 350–1
- rule
- chain 301
  - Cramer's 74
  - Leibniz 594
  - Midpoint 271
  - Open 271
  - Simpson's 271
  - Trapezoid 271, 562
- S**
- saddle point 587, 589
- salesman 505
- sample function 79–80, 235, 285
- sample point 79, 105, 285  
  even 286  
  odd 286
- sample rate 287
- sample value 79, 256, 260, 272, 286, 479
- sample vector 80, 98, 105, 183, 265, 285, 554  
  exponential 285–7, 415, 436  
  monomial 265  
  polynomial 265
- sampling 183, 297, 468
- satellite 200, 341
- savings account 476
- scalar 1, 5, 6, 12, 37, 70, 177, 389, 449, 480  
  complex 177, 476  
  real 5, 177, 476  
  unit 8  
  zero 8
- scalar multiplication 5, 8, 43, 76, 78, 87, 343, 349, 390
- scaling 341, 399, 429, 551, 561, 600  
  scaling function 399, 549, 555, 558, 559–60, 563
- scaling transformation 429
- scatter plot 469, 471, 478
- scattered 468
- scheduling 475
- Schrödinger equation 173, 394
- Schrödinger operator 437
- Schur decomposition *xii*, *xiii*, 403, 444–6
- science *vii*
- computer *vii*, *ix*, *xv*, 126, 463
  - data 1, 235
  - physical *ix*
  - social *ix*
- scientist *xiii*, *xviii*
- screen 374–5
- screw motion *xi*, 341, 373, 419, 602
- search 475, 499, 502
- segment 473
- self-adjoint 183, 399, 436, 618
- semantics 404, 467
- semi-axis 487, 498
- semi-definite
- negative 159
  - positive *xi*, 158, 161, 182, 244–5, 301, 316, 320, 433, 454, 470, 473, 514, 615, 618, 622
- semi-direct method *xii*, 475, 536, 547
- semi-magic square 104
- semi-simple *xvi*, 424
- separation of variables 227
- sequence 81, 86, 144
- serial computer 513
- serial processor 513
- serialization 475
- series 394
- exponential 596, 599, 601–2
  - Fourier 91, 191, 549, 553
  - geometric 499
  - matrix 499, 596
  - power 175
  - Taylor 87, 91
  - trigonometric 549
  - wavelet 553, 562
- sesquilinear 178–9
- set *xvii*, 80
- closed 146, 151
  - data 188, 462, 472–3
  - difference of *xvii*
  - level 585
  - open 146, 151
  - swing 335–6, 619
- sex 482

- shear *xi*, 341, 361–2, 427, 429, 600  
 shear factor 361  
 Sherman–Morrison–Woodbury formula 35  
 shift 415, 436, 531  
     phase 89, 273, 587, 609  
 shift map 415, 436  
 Shifted Inverse Power Method 526–7, 534, 539  
 sign 72  
 signal 285, 294, 549, 553  
     reconstruction 299  
     sampled 285  
 signal processing *vii*, *viii*, *x–xiii*, 1, 75, 80, 99, 102, 129, 183, 188, 235, 272, 293, 476  
 similar 73, 367, 418, 425–6, 428, 465, 498, 532, 575, 598  
 simple 120  
 simple digraph 467, 502  
 simple eigenvalue 411, 416, 493, 531, 535, 537  
 simple graph 120, 311, 463  
 simple root 411  
 simplex 339, 500  
 simplification 102  
 Simpson’s Rule 271  
 simultaneously diagonalizable 428–9  
 sinc 273  
 sine *xvii*, 176, 183, 269, 285, 581, 610  
 single precision 461  
 singular 23, 70, 314, 403, 409, 411–2, 461, 597, 631  
 singular point *xv*, 380  
 singular value *vii*, *ix*, *xiv*, 403, 454–6, 460–2, 467, 472–3, 497  
     distinct 455  
     dominant 454, 460  
     largest 460, 466, 497  
     smallest 460, 463, 466  
 Singular Value Decomposition *xii*, 403, 455, 457, 461, 473  
 singular vector 454, 461, 467, 473  
     unit 471–2  
 size 114  
 skeletal fragment 393  
 skew-adjoint 400  
 skew field 364  
 skew-symmetric 10, 47, 73, 85–6, 204, 400, 435, 439, 600–2  
 skyscraper 322–3  
 smallest eigenvalue 441  
 smooth *xviii*, 84, 279  
 Sobolev inner product 136, 144, 233  
 Sobolev norm 136  
 social science *ix*  
 soft spring 303  
 software *xvi*, 404  
 soldier 625  
 solid 236, 485, 565  
     Platonic 127  
 solid mechanics 236  
 solution *ix*, 2, 13, 21, 24, 29, 40, 63, 76, 222, 475, 484, 542, 568, 570, 577, 579, 606, 629  
     approximate 237, 541  
     complex *xiv*, 391, 575  
     constant 615  
     equilibrium 301, 405, 476, 479, 488, 493, 565, 579, 597, 622  
     exponential 381, 408  
     general 91, 107, 111, 480, 606, 618, 625  
     globally asymptotically stable 488–90, 492–3  
     homogeneous 388  
     Jordan chain 576–7, 581  
     least squares *ix*, *xi*, 237–8, 250–1, 317, 403, 458  
     linearly independent 380, 594  
     matrix 572  
     minimum norm 224, 458  
     non-resonant 628  
     nontrivial 67, 95  
     particular 107, 384, 606, 623–5, 630  
     period 2 495  
     periodic 565  
     quasi-periodic 619  
     real *xiv*, 577  
     resonant 628, 631  
     stable *viii*, 581  
     trigonometric 381  
     trivial 67  
     unbounded 586  
     unique 384, 570  
     unstable *viii*, 405, 581  
     vector-valued 592  
     zero 67, 117, 383–4, 405, 410, 476, 478, 489–90, 492–3, 581–2  
 solution space 566, 577  
 solvability *vii*  
 SOR *xii*, 475, 517–20  
 sound 624, 626  
 soundtrack 285, 293  
 source 342  
     current 301, 313, 317, 320, 629  
 space 88, 200, 341  
     column 105, 383  
     dual 350, 358, 369, 395  
     Euclidean *x*, *xi*, 75–7, 94, 99, 130, 146, 341, 403, 426, 600  
     free 394

- space (*continued*)  
 function *x*, *xiii*, *xv*, 79, 80, 83, 133, 146, 163, 185, 190, 220, 224, 301, 341, 396, 401, 541  
 Hilbert 135, 341  
 inner product 130, 140, 161, 183–4, 193, 196, 213, 219, 245, 342, 347, 350, 358  
 left null space *x*, 113  
 normed 144, 372  
 null *x*, 106  
 outer 301, 327  
 quotient 87, 105  
 row *x*, 113  
 solution 566, 577  
 three-dimensional 82–3, 88, 99, 200, 335, 373  
 vector *x*, *xi*, *xiii–xv*, *xvii*, 75–6, 82, 101, 129–30, 133, 149, 151, 177, 179, 213, 219–20, 274, 287, 301, 341–2, 349, 351, 355–7, 372, 390, 396, 401, 467, 541, 600
- space curve 283  
 space station 328, 330, 332, 339, 632  
 space-time 159–60, 235, 341, 358  
 span *ix*, *x*, *xiii*, 75, 87, 92, 95–6, 99, 100, 177, 185, 215, 217, 341  
 spark 314  
 sparse *xv*, 48, 52, 475, 536, 548  
 special function 200  
 special lower triangular *xvi*  
 special orthogonal matrix 222  
 special relativity 159, 358, 375  
 special upper triangular *xvi*  
 species 504  
 spectral condition number 525, 591  
 Spectral Decomposition 440, 598  
 spectral eigenstate 437  
 spectral energy 437  
 spectral factorization 437  
 spectral graph theory *xiv*, *xv*, 462–3  
 spectral radius *xii*, 475, 490, 492–3, 495–8, 508, 518, 520, 524  
 Spectral Theorem 437, 439, 456, 529  
 spectroscopy 608  
 spectrum *xv*, 437, 462, 467  
 graph 462, 467  
 speech 285, 404, 462, 467, 499  
 speed 102, 159, 254, 467  
 sphere 83  
 unit 83, 149–50, 363, 375, 438, 465, 473–4  
 spiral 587–8  
 spiral growth 483  
 spline *viii*, *xi*, 52, 54, 235, 263, 279, 322, 563  
*B* 284, 567  
 cardinal 284  
 periodic 280, 283  
 spline font 283  
 spline letter 283  
 spring 110, 236, 301, 303, 320, 323, 608–9, 616, 621, 623, 629  
 hard 303  
 soft 303  
 spring stiffness 303, 306, 320, 323, 327, 609, 621, 623, 629  
 square 9, 126, 167, 358, 363, 467, 505, 619  
 complete the *xi*, 129, 166, 240, 437  
 magic 104  
 non- 60, 403  
 perfect 166  
 semi-magic 104  
 unit 136  
 square grid 317, 521  
 square matrix 4, 18, 23, 31, 33, 45, 403, 416, 426, 453, 457, 495, 542, 596  
 square root *xvii*, 12, 131, 166, 171, 175, 185, 198, 214, 269, 454, 468, 611  
 matrix 439, 465, 620  
 square truss 322, 615, 632  
 squared error 255–6, 260, 272, 274  
 squares  
 least *ix*, *xi*, *xiii–xv*, 129, 132, 183, 230, 235, 237–8, 250–2, 255–6, 263, 271–2, 317, 403, 458, 468, 474  
 sum of 167  
 weighted least 252, 256, 265  
 St. Paul 501  
 stability *ix*, *xi*, *xii*, 488, 565, 590, 629  
 asymptotic 490, 493  
 Stability Theorem 577, 603  
 stabilization order 537, 540, 547, 549  
 stable 331, 405, 478, 493, 579, 581, 583–84, 586–8, 590–1, 610  
 asymptotically 405, 478, 490, 493, 579–82, 584, 586–7, 591, 597  
 globally 405, 579  
 globally asymptotically 405, 488–90, 492–3, 579, 622  
 structurally 591  
 stable eigensolution 603  
 stable equilibrium 235–6, 301–2, 579, 590, 605, 615  
 stable fixed point 493  
 stable focus 587, 589, 591  
 stable improper node 604  
 stable line 587, 589, 591  
 stable manifold 605  
 stable node 586, 588–9, 591  
 stable solution *viii*, 581  
 stable star 589–90  
 stable structure 331, 335  
 stable subspace 492, 604  
 staircase 59

- standard basis 36, 99, 111, 184, 261, 343, 349, 356, 426, 449, 450, 529
- standard deviation 468–9
- principal 472
- standard dual basis 350
- star 467, 589–90
- starting node 311
- starting vertex 121–2
- state 341, 502
- state vector 499
- static 293
- statically determinate 307, 333
- statically indeterminate 306, 315, 334
- statics 403
- station
- space 328, 330, 332, 339, 632
- statistical analysis 188, 238
- statistical fluctuation 467, 470
- statistical physics 464
- statistical test 467
- statistically independent 470
- statistics *vii–ix, xii, xiii, xv, 1, 99, 129, 132, 135, 156, 163, 238, 467, 499*
- steady-state 574
- steepest decrease 545, 583
- steepest increase 545, 582
- stiffness 303, 306, 320, 323, 327, 609, 621, 623, 629
- stiffness matrix 305, 309, 320, 327, 611, 615, 618, 622
- still image 102, 285
- stochastic
- doubly 505
- stochastic process *viii, xii, xiv, 403, 475, 499*
- stochastic system 341
- storage 102, 294, 462, 513
- Strassen's Method 51
- stress 323, 439
- stretch 341, 344, 360–2, 403, 426, 438, 440, 457, 600, 625
- principal 438–9
- strictly diagonally dominant 281, 283, 421–2, 475, 498, 510, 512, 516, 584
- strictly lower triangular *xvi, 16–8, 28, 39, 41–2, 45, 60, 85, 168, 509, 530*
- strictly upper triangular 16, 85, 509
- stroboscopic 286
- structurally stable 591
- structure *viii, xi–xiv, 236, 301, 322, 403, 565, 599, 601, 608, 610, 625, 628–32*
- atomic 203
  - mechanical 301
  - reinforced 334, 336
  - stable 331, 335
  - unstable 301, 615
- student *xviii, 504*
- subdeterminant 171
- subdiagonal 52, 492, 535
- subdominant eigenvalue 493, 502, 524, 526
- subscript 4
- subset *xvii, 79*
- closed *xvii*
  - compact 149
  - convex 150
  - thin 220
- subspace *x, xiii, 75, 82, 86, 88, 183, 213, 223, 235, 238, 245, 362, 471, 540*
- affine 87, 375, 383
- center 604
- complementary 86, 105, 217–8, 221
- complex 298, 430, 452
- conjugated 391
- dense 220
- finite-dimensional 213, 219, 248
- fundamental 114, 183, 221
- infinite-dimensional 219
- invariant *xv, 429–31, 452, 487, 492, 548, 603–4*
- Krylov *xv, 475, 536–7, 539–40, 546, 549*
- orthogonal *xv, 183, 216*
- proper 210
- real 452
- stable 492, 604
- trivial 82, 429
- unstable 604
- zero 82, 429
- substitution
- back *x, xiii, xiv, 3, 14, 21, 24, 41, 50, 53, 62, 208, 211, 282, 518*
  - forward *xiii, xiv, 3, 20, 49, 53, 282, 518*
- subtraction 48, 53, 261, 536
- suburb 501
- Successive Over–Relaxation *xii, 475, 517–20*
- sum *xvii, 86*
- absolute row 155, 496, 498, 510
  - column 10, 419, 501, 502
  - partial 554
  - row 10, 419, 502
- sum of squares 167
- sunny 499, 501
- supercomputer 1, 48
- superconducting 608, 630
- superdiagonal 52, 449, 492, 535
- superposition
- linear *vii, ix, xi, xv, 110, 222, 235, 250, 262, 342, 378, 480, 565, 630*
- Superposition Principle 75, 106, 111, 378, 388
- superscript 4

support 301, 306, 551, 562, 608–9  
     bounded 557  
 surface 83, 236, 283, 439  
 survey  
     geodetic 171  
 suspension bridge 625  
 SVD *xii*, 403, 455, 457, 461, 473  
 swing set 335, 619  
 Sylvester inequality 120  
 symmetric 157, 241  
 symmetric matrix *xi*, *xiv*, 45, 85–6, 167, 171,  
     183, 208, 216, 226, 398–9, 403, 432,  
     434, 437, 440–1, 446, 454, 465, 487,  
     532, 537, 542, 581, 585, 631  
 symmetric polynomial 417  
 symmetry *xii*, 10, 130, 133, 146, 156, 200,  
     202, 341, 358  
     conjugate 179, 184  
 symmetry analysis 599  
 system 236  
     adjoint 112, 117  
     affine 488  
     algebraic *vii*, *ix*, 341–2, 376, 386, 506, 517,  
         540  
     autonomous 403, 566, 579  
     compatible *ix*, *xi*, 8, 11, 62, 224  
     complete 572  
     complex *xiv*, 566  
     control *vii*, *xv*, 76, 99, 106, 376  
     damped 623  
     dynamical *viii*, *xii*, *xiii*, *xv*, 301–2, 396,  
         403, 407, 565, 583, 591, 603  
     electrical 183  
     elliptic 542  
     equivalent 2  
     first order 565–7, 570–2, 577, 585, 605  
     fixed point 506  
     forced 565  
     Hamiltonian 583, 585  
     higher order 605  
     homogeneous *vii*, *xi*, *xii*, 67, 95, 99, 106,  
         108, 342, 376, 378, 384, 388, 394, 409,  
         571, 585, 592  
     ill-conditioned 57, 211, 461  
     implicit 492  
     incompatible *xi*, 62  
     inhomogeneous *vii*, *xi*, *xii*, 67, 106, 110–1,  
         342, 376, 383–4, 388, 394, 565, 585,  
         605–6, 630  
     inconsistent 62  
     iterative 53, 475, 481, 488, 492–3, 563  
     large 475  
     linear *vii*, *ix*, *xi*, 4, 6, 20, 23, 40, 59, 63, 67,  
         75, 99, 105–7, 376, 461, 475, 541, 565,  
         571, 577

system (*continued*)  
     linear algebraic *vii*, *ix*, 341–2, 376, 386,  
         506, 517, 540  
     lower triangular 3, 20  
     Newtonian 614  
     non-autonomous 570, 598  
     nonlinear 64, 66, 342, 475, 568, 604  
     order of 481, 493  
     orthogonal 552  
     planar 565, 585  
     singular 461  
     sparse *xv*, 52, 475, 536  
     stochastic 341  
     triangular 2, 20, 29, 197, 542  
     trivial 590–1  
     two-dimensional 585  
     undamped 623, 630–1  
     unforced 622  
     weak 132, 398, 540–1, 546  
 system of ordinary differential equations  
     *ix*, *xii*–*xv*, 342, 530, 566, 584, 571, 579,  
         592, 608, 630  
     second order *xii*, 618

## T

Tacoma Narrows Bridge 626  
 tangent *xvii*, 341, 600  
 target *xvi*, 342  
 taxi 501  
 Taylor polynomial 269, 324, 383  
 Taylor series 87, 91  
 technology 555  
 temperature 258  
 tension 327  
 tensor 4  
     inertia 439  
 terminal 317  
 terminating node 311, 322  
 test  
     statistical 467  
 tetrahedron 127, 139, 321, 619–20  
 theater 293  
 theorem *viii*  
     Cayley–Hamilton 420, 453  
     Center Manifold 604  
     Fundamental, of Algebra 98, 124, 415  
     Fundamental, of Calculus 347, 356, 606  
     Fundamental, of Linear Algebra 114, 461  
     Fundamental, of Principal Component  
         Analysis 472  
     Gershgorin Circle 420, 475, 503  
     Jordan Basis 448, 450  
     Perron–Frobenius 501  
     Pythagorean 130–2  
     Rolle’s 231

- theorem (*continued*)  
 Spectral 437, 439, 456, 529  
 Stability 577, 603  
 Weierstrass Approximation 220
- theory *viii, xiii*  
 category *viii*  
 control 235  
 function 135  
 graph *viii, x, xiv, 12*  
 group *xii, 464, 599*  
 measure 135  
 potential 173
- thermodynamics 183, 236, 381, 403
- thin 220
- three-dimensional space 82–3, 88, 99, 200, 335, 373
- time *viii, 475–6, 568*  
 initial 621  
 space- 159–60, 235, 341, 358
- time reversal 569
- topology 120, 146, 151, 312
- total variance 472
- tower 322
- trace 10, 85, 170–1, 415, 417, 586, 590–1, 596, 600
- trajectory 568, 600, 602, 616
- transform  
 Cayley 204  
 discrete Fourier *xi, xv, 183, 272, 285, 289, 295*  
 fast Fourier *xi, 235, 296*  
 Fourier 376, 559  
 integral 376  
 Laplace 376  
 wavelet transform 554
- transformation 342, 358
- affine *ix, xii, xiv, 341, 370–3, 377, 419, 603*  
 identity 348, 429  
 linear *ix, xiii, xiv, 341–2, 358, 403, 426, 429, 457, 554, 599*  
 scaling 429  
 self-adjoint 436  
 shearing 361
- transient 627
- transition matrix 4499–501, 505, 525, 528, 536, 598  
 regular 501
- transitional probability 501–2
- translation *xi, 328, 331, 341, 346, 370–2, 419, 550–1, 556–8, 561–2, 601, 616*
- transmission 272, 294
- transpose *x, xii, 43–5, 72, 112, 114, 162, 200, 222, 304, 314, 342, 351, 357, 369, 395, 397, 399, 416, 422, 456, 502, 597*  
 Hermitian 205, 444
- trapezoid 126
- Trapezoid Rule 271, 562
- travel company 258
- traveling salesman 505
- tree 127, 322, 467
- tribonacci number 486
- triangle 125, 363, 371, 467, 505, 632  
 equilateral 328, 500, 619
- triangle inequality 129, 142–4, 146, 154, 179, 498
- triangular  
 block upper 74, 535  
 lower *xvi, 3, 16–7, 20, 39, 73, 518*  
 positive upper 205, 529–30  
 special *xvi*  
 strictly lower *xvi, 16–8, 28, 39, 41–2, 45, 60, 85, 168, 509, 530*  
 strictly upper 16, 85, 509  
 upper *xvi, 13, 16, 20, 23–4, 28, 39, 49, 70, 204–5, 210, 425, 428, 444–6, 465, 518, 527, 530, 532, 602*
- triangular form 2, 14
- triangular matrix  
 lower *xvi, 16–7, 20, 39, 73, 518*  
 upper *xvi, 13, 16, 23–4, 28, 39, 70, 204–5, 210, 425, 428, 444–6, 465, 518, 527, 530, 532, 602*
- triangular system 2, 20, 29, 197, 542
- triangularize 444
- triatomic molecule 616, 619, 632
- tricirculant matrix 54, 282–3, 420, 436
- tridiagonal matrix 52, 281, 304, 419, 492, 512, 526, 532, 535–6, 539, 542
- tridiagonal solution algorithm 52, 282
- tridiagonalization 532, 535
- trigonometric ansatz 609, 618, 623, 625, 630
- trigonometric approximation 271, 273
- trigonometric function *xi, xiv, xvii, 89, 164, 175–6, 183, 235, 272, 292, 578, 580–1*  
 complex 176–7
- trigonometric identity 175
- trigonometric integral 175, 177, 624
- trigonometric interpolation 86, 235, 287, 293
- trigonometric monomial 190
- trigonometric polynomial 75, 90–1, 94, 176, 190, 273
- trigonometric series 549
- trigonometric solution 381
- trivial solution 67
- trivial subspace 82, 429
- trivial system 590–1
- truss 322, 615, 632
- tuning fork 624
- turkey 258
- two-dimensional system 585
- 2-norm 145

typography 283

## U

unbiased 468  
 unbounded 581, 586, 603  
 Uncertainty Principle 355  
 uncorrelated 469, 472  
 undamped 623, 630–1  
 underdamped 621, 627, 629  
 underflow 524  
 undergraduate 9  
 under-relaxed 518  
 underwater vehicle 200  
 undetermined coefficients 372, 385–6, 500, 623  
 unforced 622  
 uniform convergence 562  
 uniform distribution 468  
 union *xvii*, 86  
 unipotent 16–7  
 uniqueness 1, 23, 40, 380, 383–4, 401, 479, 568, 570, 593, 610  
 unit 76  
 additive 7  
 imaginary *xvii*, 173  
 multiplicative 7  
 unit ball 85, 149–50, 473  
 unit circle *xvii*, 132, 288, 442, 530  
 unit cross polytope 149  
 unit diamond 149  
 unit disk 136, 371, 503  
 unit eigenvector 471, 493, 496  
 unit element 148  
 unit function 148  
 unit octahedron 149  
 unit polynomial 148  
 unit radius *viii*  
 unit scalar 8  
 unit sphere 83, 149–50, 363, 375, 438, 465, 473–4  
 unit square 136  
 unit vector 141, 148, 150, 184, 208, 325, 327, 441, 443, 524, 532, 576, 602  
 unitary matrix 205, 212, 439, 444–6  
 United States 259  
 unitriangular *xvi*, 16  
 lower *xvi*, 16–8, 20, 28, 39, 41–3, 45, 60, 85, 168, 530  
 upper *xvi*, 16, 18, 38, 41–3, 543  
 unity  
 root of 288, 292, 296  
 universe *vii*, 160, 358, 403, 628  
 unknown 4, 6, 506  
 unstable 314, 405, 478, 581, 583–4, 586, 591, 619

unstable eigensolution 603  
 unstable equilibrium 235–6, 301, 590  
 unstable focus 588–9, 591  
 unstable improper node 588–9, 591  
 unstable line 587, 589  
 unstable manifold 605  
 unstable mode 615, 618–9  
 unstable node 587–9, 591  
 unstable solution *viii*, 405, 581  
 unstable star 589–90  
 unstable structure 301, 615  
 unstable subspace 604  
 upper Hessenberg matrix 535–6, 539, 542  
 upper triangular *xvi*, 13, 16, 23–4, 28, 39, 70  
 block 74, 535  
 positive 205, 529–30  
 special *xvi*  
 strictly 16, 85, 509  
 upper triangular system 20, 49  
 upper unitriangular *xvi*, 16, 18, 38, 41–3, 543  
 uranium 404

## V

valley 236  
 value  
 absolute *xvii*  
 boundary *vii*, *x*, *xi*, *xv*, 54, 75, 92, 99, 136, 183, 222, 235, 322, 342, 376–6, 386, 389, 397, 399, 541–2  
 characteristic 408  
 expected 468  
 initial 376, 386, 570, 594, 598, 606  
 proper 408  
 sample 79, 256, 260, 272, 286, 479  
 singular *vii*, *ix*, *xii*, *xiv*, 403, 454–7, 460–2, 466–7, 472–3, 497  
 Vandermonde matrix 20, 74, 260, 268  
 variable *viii*, 2, 62, 605  
 basic 62–3, 118  
 change of 172, 232, 234  
 complex 172  
 free 62, 63, 67, 96, 108, 119–20, 315  
 phase plane 567  
 separation of 227  
 variance 468–71, 473  
 principal 472–3  
 total 472  
 unbiased 468  
 variation *xii*, 235  
 variation of parameters 385, 606, 623  
 vat 622  
 vector *ix*, *x*, *xiii*, *xiv*, *xvii*, 1, 75, 129, 223, 341, 457, 480, 571, 578  
 Arnoldi 538–40, 542, 547  
 battery 317

- vector (*continued*)
   
characteristic 408
   
circuit 312
   
column 4, 6, 46, 48, 77, 130, 350–1
   
complex 129, 433
   
constant 588
   
current source 314, 318, 320
   
data 254
   
displacement 302, 312, 325, 620
   
elongation 302, 325
   
error 254, 508, 514
   
force 327
   
gradient 349, 545, 582
   
Gram–Schmidt 194
   
Householder 210–1, 536
   
image 473
   
initial 475, 540
   
Krylov 537, 547
   
measurement 468, 470
   
mechanism 336
   
nonzero 95
   
normal 217
   
normalized 468, 470
   
orthogonal *xiii–xv*, 140, 185
   
parallel 142, 147
   
probability 473, 500–1
   
proper 408
   
real 76, 391
   
residual 237, 522, 541, 544–5, 548
   
row 4, 6, 130, 350–1
   
sample 80, 98, 105, 183, 265, 285–7, 436, 554
   
singular 454, 461, 467, 473
   
standard basis 36, 99, 111, 184, 261, 343, 349, 356, 426, 449, 450, 529
   
state 499
   
unit 141, 148, 150, 184, 208, 325, 327, 441, 443, 524, 532, 576, 602
   
velocity 574
   
voltage 312, 320
   
zero 7, 67, 77, 82, 93, 131, 407, 493
- vector addition 77, 82, 390
- vector calculus 353, 365
- vector field 81, 574
- vector product 6, 130
- vector space *x*, *xi*, *xiii*, *xvii*, 75–6, 82, 130, 341, 349, 600
   
complex *xi*, *xiv*, 76, 129, 177, 179, 287, 342, 390
   
conjugated 390
   
finite-dimensional *xiii*, *xiv*, 101, 149, 356
   
high-dimensional 467
   
infinite-dimensional *x*, *xiv*, *xv*, 101, 129, 133, 149, 151, 213, 219–20, 274, 301, 341, 349, 351, 355–6, 396, 401, 541
- vector space (*continued*)
   
isomorphic 356
   
normed 144, 372
   
quotient 87, 105, 357
   
real *xi*, 76, 342
   
vector-valued function 80, 98, 136, 341, 605
   
vector-valued solution 592
   
vehicle 200, 254, 616
   
velocity *viii*, 80, 254, 259, 365, 615, 618, 620–1
   
initial 609–10, 618, 622
   
velocity vector field 574
   
vertex 120, 122, 125–6, 311, 463, 502, 505
   
ending 122
   
starting 121–2
   
vibration *viii*, *ix*, *xii*, 399, 565, 608–9, 611, 621–2, 625–6, 629–31
   
internal 565, 624
   
mechanical 565
   
natural 614
   
periodic 618
   
quasi-periodic 565, 617–8
   
resonant 565, 625
   
vibrational force 630
   
vibrational frequency 610, 621, 624
   
vibrational mode 616
   
video 48, 102, 188, 200, 285, 294–5, 341
   
vision
   
  computer 375, 499
   
voltage 121, 311–2, 319–20, 626–7
   
Voltage Balance Law 312, 314, 629
   
voltage potential 311–2, 315, 318, 320
   
voltage vector 312, 320
   
Volterra integral equation 378
   
volume 465
- W**
- walk
   
  random 463
   
wall 322
   
waste 258, 406
   
water 583
   
water molecule 620, 626, 632
   
wave
   
  electromagnetic 626
   
wave function 173, 341
   
wavelet *xii*, *xiv*, *xv*, 102, 227, 476, 549, 552, 555, 560
   
  Daubechies 555, 562
   
  daughter 550, 556, 563
   
  Haar 549–50, 552–3, 555, 562
   
  mother 550, 552, 555–6, 558
   
wavelet basis 102, 189, 204, 283, 550, 552, 555–6, 562
   
wavelet coefficient 470

wavelet matrix 204, 224, 552, 554  
wavelet series 553, 562  
wavelet transform 554  
weak formulation 132, 398, 540–1, 546  
weather 48, 407, 499, 501  
web page 126, 463, 502  
Weierstrass Approximation Theorem 220  
weight 132, 135, 256, 502  
    atomic 620  
weight matrix 256  
weighted adjoint 397  
weighted angle 138  
weighted digraph 311, 502  
weighted Gram matrix 163, 247  
weighted graph 311  
weighted inner product 131, 135, 182, 246,  
    265, 309, 396, 435, 543, 618  
weighted integral 182  
weighted least squares 252, 256, 265  
weighted norm 131, 135, 237, 252, 468  
weighted normal equation 247, 252, 256, 317  
wheel 287  
white collar 504  
wind 323  
wind instrument 626  
wire 120, 301, 311–3, 317, 319, 327  
withdrawal 476  
Wronskian 98

## Y

year 475–6  
Young matrix 519–20, 579  
YouTube 626

## Z

zero *xvi*  
    mean 10, 84, 468, 470  
zero column 59  
zero correlation 470  
zero determinant 70  
zero eigenspace 434  
zero eigenvalue 412, 421, 433–4, 581, 615  
zero element 76, 79, 82, 87, 140, 342  
zero entry 52, 59, 449, 501  
zero function 79, 83, 134, 343, 405  
zero map 361  
zero matrix 7, 8, 61, 77, 361, 457, 488, 597  
zero polynomial 78  
zero potential 315  
zero row 70, 73  
zero scalar 8  
zero solution 67, 117, 383–4, 405, 410, 476,  
    478, 489–90, 492–3, 581–2  
zero subspace 82, 429  
zero vector 7, 67, 77, 82, 93, 131, 407, 493