Matrix Analysis and Applied Linear Algebra

Carl D. Meyer

siam

Contents

Preface ix

1. Linear Equations **1**

1.1 Introduction 1
1.2 Gaussian Elimination and Matrices 3
1.3 Gauss–Jordan Method 15
1.4 Two-Point Boundary Value Problems 18
1.5 Making Gaussian Elimination Work 21
1.6 Ill-Conditioned Systems 33

2. Rectangular Systems and Echelon Forms . . . **41**

2.1 Row Echelon Form and Rank 41
2.2 Reduced Row Echelon Form 47
2.3 Consistency of Linear Systems 53
2.4 Homogeneous Systems 57
2.5 Nonhomogeneous Systems 64
2.6 Electrical Circuits 73

3. Matrix Algebra **79**

3.1 From Ancient China to Arthur Cayley 79
3.2 Addition and Transposition 81
3.3 Linearity 89
3.4 Why Do It This Way 93
3.5 Matrix Multiplication 95
3.6 Properties of Matrix Multiplication 105
3.7 Matrix Inversion 115
3.8 Inverses of Sums and Sensitivity 124
3.9 Elementary Matrices and Equivalence 131
3.10 The LU Factorization 141

4. Vector Spaces **159**

4.1 Spaces and Subspaces 159
4.2 Four Fundamental Subspaces 169
4.3 Linear Independence 181
4.4 Basis and Dimension 194

4.5 More about Rank 210
4.6 Classical Least Squares 223
4.7 Linear Transformations 238
4.8 Change of Basis and Similarity 251
4.9 Invariant Subspaces 259

5. Norms, Inner Products, and Orthogonality . . 269
5.1 Vector Norms 269
5.2 Matrix Norms 279
5.3 Inner-Product Spaces 286
5.4 Orthogonal Vectors 294
5.5 Gram–Schmidt Procedure 307
5.6 Unitary and Orthogonal Matrices 320
5.7 Orthogonal Reduction 341
5.8 Discrete Fourier Transform 356
5.9 Complementary Subspaces 383
5.10 Range-Nullspace Decomposition 394
5.11 Orthogonal Decomposition 403
5.12 Singular Value Decomposition 411
5.13 Orthogonal Projection 429
5.14 Why Least Squares? 446
5.15 Angles between Subspaces 450

6. Determinants 459
6.1 Determinants 459
6.2 Additional Properties of Determinants 475

7. Eigenvalues and Eigenvectors 489
7.1 Elementary Properties of Eigensystems 489
7.2 Diagonalization by Similarity Transformations . . 505
7.3 Functions of Diagonalizable Matrices 525
7.4 Systems of Differential Equations 541
7.5 Normal Matrices 547
7.6 Positive Definite Matrices 558
7.7 Nilpotent Matrices and Jordan Structure 574
7.8 Jordan Form 587
7.9 Functions of Nondiagonalizable Matrices 599

7.10 Difference Equations, Limits, and Summability . . 616
7.11 Minimum Polynomials and Krylov Methods . . . 642

8. Perron–Frobenius Theory **661**
8.1 Introduction 661
8.2 Positive Matrices 663
8.3 Nonnegative Matrices 670
8.4 Stochastic Matrices and Markov Chains 687

Index 705

Preface

Scaffolding

Reacting to criticism concerning the lack of motivation in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolding in place after the construction has been completed. His philosophy epitomized the formal presentation and teaching of mathematics throughout the nineteenth and twentieth centuries, and it is still commonly found in mid-to-upper-level mathematics textbooks. The inherent efficiency and natural beauty of mathematics are compromised by straying too far from Gauss's viewpoint. But, as with most things in life, appreciation is generally preceded by some understanding seasoned with a bit of maturity, and in mathematics this comes from seeing some of the scaffolding.

Purpose, Gap, and Challenge

The purpose of this text is to present the contemporary theory and applications of linear algebra to university students studying mathematics, engineering, or applied science at the postcalculus level. Because linear algebra is usually encountered between basic problem solving courses such as calculus or differential equations and more advanced courses that require students to cope with mathematical rigors, the challenge in teaching applied linear algebra is to expose some of the scaffolding while conditioning students to appreciate the utility and beauty of the subject. Effectively meeting this challenge and bridging the inherent gaps between basic and more advanced mathematics are primary goals of this book.

Rigor and Formalism

To reveal portions of the scaffolding, narratives, examples, and summaries are used in place of the formal definition–theorem–proof development. But while well-chosen examples can be more effective in promoting understanding than rigorous proofs, and while precious classroom minutes cannot be squandered on theoretical details, I believe that all scientifically oriented students should be exposed to some degree of mathematical thought, logic, and rigor. And if logic and rigor are to reside anywhere, they have to be in the textbook. So even when logic and rigor are not the primary thrust, they are always available. Formal definition–theorem–proof designations are not used, but definitions, theorems, and proofs nevertheless exist, and they become evident as a student's maturity increases. A significant effort is made to present a linear development that avoids forward references, circular arguments, and dependence on prior knowledge of the subject. This results in some inefficiencies—e.g., the matrix 2-norm is presented

before eigenvalues or singular values are thoroughly discussed. To compensate, I try to provide enough "wiggle room" so that an instructor can temper the inefficiencies by tailoring the approach to the students' prior background.

Comprehensiveness and Flexibility

A rather comprehensive treatment of linear algebra and its applications is presented and, consequently, the book is not meant to be devoured cover-to-cover in a typical one-semester course. However, the presentation is structured to provide flexibility in topic selection so that the text can be easily adapted to meet the demands of different course outlines without suffering breaks in continuity. Each section contains basic material paired with straightforward explanations, examples, and exercises. But every section also contains a degree of depth coupled with thought-provoking examples and exercises that can take interested students to a higher level. The exercises are formulated not only to make a student think about material from a current section, but they are designed also to pave the way for ideas in future sections in a smooth and often transparent manner. The text accommodates a variety of presentation levels by allowing instructors to select sections, discussions, examples, and exercises of appropriate sophistication. For example, traditional one-semester undergraduate courses can be taught from the basic material in Chapter 1 (Linear Equations); Chapter 2 (Rectangular Systems and Echelon Forms); Chapter 3 (Matrix Algebra); Chapter 4 (Vector Spaces); Chapter 5 (Norms, Inner Products, and Orthogonality); Chapter 6 (Determinants); and Chapter 7 (Eigenvalues and Eigenvectors). The level of the course and the degree of rigor are controlled by the selection and depth of coverage in the latter sections of Chapters 4, 5, and 7. An upper-level course might consist of a quick review of Chapters 1, 2, and 3 followed by a more in-depth treatment of Chapters 4, 5, and 7. For courses containing advanced undergraduate or graduate students, the focus can be on material in the latter sections of Chapters 4, 5, 7, and Chapter 8 (Perron–Frobenius Theory of Nonnegative Matrices). A rich two-semester course can be taught by using the text in its entirety.

What Does "Applied" Mean?

Most people agree that linear algebra is at the heart of applied science, but there are divergent views concerning what "applied linear algebra" really means; the academician's perspective is not always the same as that of the practitioner. In a poll conducted by SIAM in preparation for one of the triannual SIAM conferences on applied linear algebra, a diverse group of internationally recognized scientific corporations and government laboratories was asked how linear algebra finds application in their missions. The overwhelming response was that the primary use of linear algebra in applied industrial and laboratory work involves the development, analysis, and implementation of numerical algorithms along with some discrete and statistical modeling. The applications in this book tend to reflect this realization. While most of the popular "academic" applications are included, and "applications" to other areas of mathematics are honestly treated,

there is an emphasis on numerical issues designed to prepare students to use linear algebra in scientific environments outside the classroom.

Computing Projects

Computing projects help solidify concepts, and I include many exercises that can be incorporated into a laboratory setting. But my goal is to write a mathematics text that can last, so I don't muddy the development by marrying the material to a particular computer package or language. I am old enough to remember what happened to the FORTRAN- and APL-based calculus and linear algebra texts that came to market in the 1970s. I provide instructors with a flexible environment that allows for an ancillary computing laboratory in which any number of popular packages and lab manuals can be used in conjunction with the material in the text.

History

Finally, I believe that revealing only the scaffolding without teaching something about the scientific architects who erected it deprives students of an important part of their mathematical heritage. It also tends to dehumanize mathematics, which is the epitome of human endeavor. Consequently, I make an effort to say things (sometimes very human things that are not always complimentary) about the lives of the people who contributed to the development and applications of linear algebra. But, as I came to realize, this is a perilous task because writing history is frequently an interpretation of facts rather than a statement of facts. I considered documenting the sources of the historical remarks to help mitigate the inevitable challenges, but it soon became apparent that the sheer volume required to do so would skew the direction and flavor of the text. I can only assure the reader that I made an effort to be as honest as possible, and I tried to corroborate "facts." Nevertheless, there were times when interpretations had to be made, and these were no doubt influenced by my own views and experiences.

Supplements

Included with this text is a solutions manual and a CD-ROM. The solutions manual contains the solutions for each exercise given in the book. The solutions are constructed to be an integral part of the learning process. Rather than just providing answers, the solutions often contain details and discussions that are intended to stimulate thought and motivate material in the following sections. The CD, produced by Vickie Kearn and the people at SIAM, contains the entire book along with the solutions manual in PDF format. This electronic version of the text is completely searchable and linked. With a click of the mouse a student can jump to a referenced page, equation, theorem, definition, or proof, and then jump back to the sentence containing the reference, thereby making learning quite efficient. In addition, the CD contains material that extends historical remarks in the book and brings them to life with a large selection of

portraits, pictures, attractive graphics, and additional anecdotes. The supporting Internet site at `MatrixAnalysis.com` contains updates, errata, new material, and additional supplements as they become available.

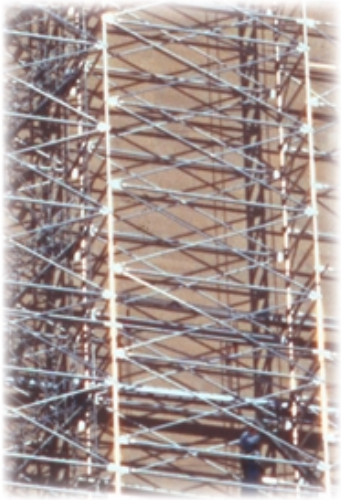
SIAM

I thank the SIAM organization and the people who constitute it (the infrastructure as well as the general membership) for allowing me the honor of publishing my book under their name. I am dedicated to the goals, philosophy, and ideals of SIAM, and there is no other company or organization in the world that I would rather have publish this book. In particular, I am most thankful to Vickie Kearn, publisher at SIAM, for the confidence, vision, and dedication she has continually provided, and I am grateful for her patience that allowed me to write the book that I wanted to write. The talented people on the SIAM staff went far above and beyond the call of ordinary duty to make this project special. This group includes Lois Sellers (art and cover design), Michelle Montgomery and Kathleen LeBlanc (promotion and marketing), Marianne Will and Deborah Poulson (copy for CD-ROM biographies), Laura Helfrich and David Comdico (design and layout of the CD-ROM), Kelly Cuomo (linking the CD-ROM), and Kelly Thomas (managing editor for the book). Special thanks goes to Jean Anderson for her eagle-sharp editor's eye.

Acknowledgments

This book evolved over a period of several years through many different courses populated by hundreds of undergraduate and graduate students. To all my students and colleagues who have offered suggestions, corrections, criticisms, or just moral support, I offer my heartfelt thanks, and I hope to see as many of you as possible at some point in the future so that I can convey my feelings to you in person. I am particularly indebted to Michele Benzi for conversations and suggestions that led to several improvements. All writers are influenced by people who have written before them, and for me these writers include (in no particular order) Gil Strang, Jim Ortega, Charlie Van Loan, Leonid Mirsky, Ben Noble, Pete Stewart, Gene Golub, Charlie Johnson, Roger Horn, Peter Lancaster, Paul Halmos, Franz Hohn, Nick Rose, and Richard Bellman—thanks for lighting the path. I want to offer particular thanks to Richard J. Painter and Franklin A. Graybill, two exceptionally fine teachers, for giving a rough Colorado farm boy a chance to pursue his dreams. Finally, neither this book nor anything else I have done in my career would have been possible without the love, help, and unwavering support from Bethany, my friend, partner, and wife. Her multiple readings of the manuscript and suggestions were invaluable. I dedicate this book to Bethany and our children, Martin and Holly, to our granddaughter, Margaret, and to the memory of my parents, Carl and Louise Meyer.

Carl D. Meyer
April 19, 2000



CHAPTER 1

Linear Equations

1.1 INTRODUCTION

A fundamental problem that surfaces in all mathematical sciences is that of analyzing and solving m algebraic equations in n unknowns. The study of a system of simultaneous linear equations is in a natural and indivisible alliance with the study of the rectangular array of numbers defined by the coefficients of the equations. This link seems to have been made at the outset.

The earliest recorded analysis of simultaneous equations is found in the ancient Chinese book *Chiu-chang Suan-shu* (*Nine Chapters on Arithmetic*), estimated to have been written some time around 200 B.C. In the beginning of Chapter VIII, there appears a problem of the following form.

> *Three sheafs of a good crop, two sheafs of a mediocre crop, and one sheaf of a bad crop are sold for* 39 *dou. Two sheafs of good, three mediocre, and one bad are sold for* 34 *dou; and one good, two mediocre, and three bad are sold for* 26 *dou. What is the price received for each sheaf of a good crop, each sheaf of a mediocre crop, and each sheaf of a bad crop?*

Today, this problem would be formulated as three equations in three unknowns by writing

$$\begin{aligned} 3x + 2y + z &= 39, \\ 2x + 3y + z &= 34, \\ x + 2y + 3z &= 26, \end{aligned}$$

where x, y, and z represent the price for one sheaf of a good, mediocre, and bad crop, respectively. The Chinese saw right to the heart of the matter. They placed the coefficients (represented by colored bamboo rods) of this system in

a square array on a "counting board" and then manipulated the lines of the array according to prescribed rules of thumb. Their counting board techniques and rules of thumb found their way to Japan and eventually appeared in Europe with the colored rods having been replaced by numerals and the counting board replaced by pen and paper. In Europe, the technique became known as *Gaussian elimination* in honor of the German mathematician Carl Gauss,[1] whose extensive use of it popularized the method.

Because this elimination technique is fundamental, we begin the study of our subject by learning how to apply this method in order to compute solutions for linear equations. After the computational aspects have been mastered, we will turn to the more theoretical facets surrounding linear systems.

[1] Carl Friedrich Gauss (1777–1855) is considered by many to have been the greatest mathematician who has ever lived, and his astounding career requires several volumes to document. He was referred to by his peers as the "prince of mathematicians." Upon Gauss's death one of them wrote that "His mind penetrated into the deepest secrets of numbers, space, and nature; He measured the course of the stars, the form and forces of the Earth; He carried within himself the evolution of mathematical sciences of a coming century." History has proven this remark to be true.

1.2 GAUSSIAN ELIMINATION AND MATRICES

The problem is to calculate, if possible, a common solution for a system of m linear algebraic equations in n unknowns

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m, \end{array}$$

where the x_i 's are the unknowns and the a_{ij} 's and the b_i 's are known constants. The a_{ij} 's are called the ***coefficients*** of the system, and the set of b_i 's is referred to as the ***right-hand side*** of the system. For any such system, there are exactly three possibilities for the set of solutions.

Three Possibilities

- **UNIQUE SOLUTION:** There is one and only one set of values for the x_i 's that satisfies all equations simultaneously.
- **NO SOLUTION:** There is no set of values for the x_i 's that satisfies all equations simultaneously—the solution set is empty.
- **INFINITELY MANY SOLUTIONS:** There are infinitely many different sets of values for the x_i 's that satisfy all equations simultaneously. It is not difficult to prove that if a system has more than one solution, then it has infinitely many solutions. For example, it is impossible for a system to have exactly two different solutions.

Part of the job in dealing with a linear system is to decide which one of these three possibilities is true. The other part of the task is to compute the solution if it is unique or to describe the set of all solutions if there are many solutions. Gaussian elimination is a tool that can be used to accomplish all of these goals.

Gaussian elimination is a methodical process of systematically transforming one system into another simpler, but equivalent, system (two systems are called ***equivalent*** if they possess equal solution sets) by successively eliminating unknowns and eventually arriving at a system that is easily solvable. The elimination process relies on three simple operations by which to transform one system to another equivalent system. To describe these operations, let E_k denote the kth equation

$$E_k: \quad a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n = b_k$$

and write the system as

$$\mathcal{S} = \left\{ \begin{array}{c} E_1 \\ E_2 \\ \vdots \\ E_m \end{array} \right\}.$$

For a linear system $\mathcal{S}$, each of the following three ***elementary operations*** results in an equivalent system $\mathcal{S}'$.

(1) Interchange the ith and j^{th} equations. That is, if

$$\mathcal{S} = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_i \\ \vdots \\ E_j \\ \vdots \\ E_m \end{array} \right\}, \quad \text{then} \quad \mathcal{S}' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_j \\ \vdots \\ E_i \\ \vdots \\ E_m \end{array} \right\}. \tag{1.2.1}$$

(2) Replace the ith equation by a nonzero multiple of itself. That is,

$$\mathcal{S}' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ \alpha E_i \\ \vdots \\ E_m \end{array} \right\}, \quad \text{where} \quad \alpha \neq 0. \tag{1.2.2}$$

(3) Replace the jth equation by a combination of itself plus a multiple of the ith equation. That is,

$$\mathcal{S}' = \left\{ \begin{array}{c} E_1 \\ \vdots \\ E_i \\ \vdots \\ E_j + \alpha E_i \\ \vdots \\ E_m \end{array} \right\}. \tag{1.2.3}$$

Providing explanations for why each of these operations cannot change the solution set is left as an exercise.

The most common problem encountered in practice is the one in which there are n equations as well as n unknowns—called a ***square system***—for which there is a unique solution. Since Gaussian elimination is straightforward for this case, we begin here and later discuss the other possibilities. What follows is a detailed description of Gaussian elimination as applied to the following simple (but typical) square system:

$$\begin{aligned} 2x + y + z &= 1, \\ 6x + 2y + z &= -1, \\ -2x + 2y + z &= 7. \end{aligned} \tag{1.2.4}$$

At each step, the strategy is to focus on one position, called the ***pivot position,*** and to eliminate all terms below this position using the three elementary operations. The coefficient in the pivot position is called a ***pivotal element*** (or simply a ***pivot***), while the equation in which the pivot lies is referred to as the ***pivotal equation.*** Only nonzero numbers are allowed to be pivots. If a coefficient in a pivot position is ever 0, then the pivotal equation is interchanged with an equation *below* the pivotal equation to produce a nonzero pivot. (This is always possible for square systems possessing a unique solution.) Unless it is 0, the first coefficient of the first equation is taken as the first pivot. For example, the circled 2 in the system below is the pivot for the first step:

$$\begin{aligned} 2x + y + z &= 1, \\ 6x + 2y + z &= -1, \\ -2x + 2y + z &= 7. \end{aligned}$$

Step 1. Eliminate all terms below the first pivot.

- Subtract three times the first equation from the second so as to produce the equivalent system:

$$\begin{aligned} 2x + y + z &= 1, \\ - y - 2z &= -4 \qquad (E_2 - 3E_1), \\ -2x + 2y + z &= 7. \end{aligned}$$

- Add the first equation to the third equation to produce the equivalent system:

$$\begin{aligned} 2x + y + z &= 1, \\ - y - 2z &= -4, \\ 3y + 2z &= 8 \qquad (E_3 + E_1). \end{aligned}$$

Step 2. Select a new pivot.

- For the time being, select a new pivot by moving down and to the right.[2] If this coefficient is not 0, then it is the next pivot. Otherwise, interchange with an equation *below* this position so as to bring a nonzero number into this pivotal position. In our example, -1 is the second pivot as identified below:

$$\begin{aligned} 2x + y + z &= 1, \\ -1y - 2z &= -4, \\ 3y + 2z &= 8. \end{aligned}$$

Step 3. Eliminate all terms below the second pivot.

- Add three times the second equation to the third equation so as to produce the equivalent system:

$$\begin{aligned} 2x + y + z &= 1, \\ -1y - 2z &= -4, \\ -4z &= -4 \qquad (E_3 + 3E_2). \end{aligned} \tag{1.2.5}$$

- In general, at each step you move down and to the right to select the next pivot, then eliminate all terms below the pivot until you can no longer proceed. In this example, the third pivot is -4, but since there is nothing below the third pivot to eliminate, the process is complete.

At this point, we say that the system has been ***triangularized.*** A triangular system is easily solved by a simple method known as ***back substitution*** in which the last equation is solved for the value of the last unknown and then substituted back into the penultimate equation, which is in turn solved for the penultimate unknown, etc., until each unknown has been determined. For our example, solve the last equation in (1.2.5) to obtain

$$z = 1.$$

Substitute $z = 1$ back into the second equation in (1.2.5) and determine

$$y = 4 - 2z = 4 - 2(1) = 2.$$

[2] The strategy of selecting pivots in numerical computation is usually a bit more complicated than simply using the next coefficient that is down and to the right. Use the down-and-right strategy for now, and later more practical strategies will be discussed.

Finally, substitute $z = 1$ and $y = 2$ back into the first equation in (1.2.5) to get

$$x = \frac{1}{2}(1 - y - z) = \frac{1}{2}(1 - 2 - 1) = -1,$$

which completes the solution.

It should be clear that there is no reason to write down the symbols such as "x," "y," "z," and "$=$" at each step since we are only manipulating the coefficients. If such symbols are discarded, then a system of linear equations reduces to a rectangular array of numbers in which each horizontal line represents one equation. For example, the system in (1.2.4) reduces to the following array:

$$\left(\begin{array}{rrr|r} 2 & 1 & 1 & 1 \\ 6 & 2 & 1 & -1 \\ -2 & 2 & 1 & 7 \end{array}\right). \qquad \text{(The line emphasizes where = appeared.)}$$

The array of coefficients—the numbers on the left-hand side of the vertical line—is called the ***coefficient matrix*** for the system. The entire array—the coefficient matrix augmented by the numbers from the right-hand side of the system—is called the ***augmented matrix*** associated with the system. If the coefficient matrix is denoted by $\mathbf{A}$ and the right-hand side is denoted by $\mathbf{b}$, then the augmented matrix associated with the system is denoted by $[\mathbf{A}|\mathbf{b}]$.

Formally, a ***scalar*** is either a real number or a complex number, and a ***matrix*** is a rectangular array of scalars. It is common practice to use uppercase boldface letters to denote matrices and to use the corresponding lowercase letters with two subscripts to denote individual entries in a matrix. For example,

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

The first subscript on an individual entry in a matrix designates the ***row*** (the horizontal line), and the second subscript denotes the ***column*** (the vertical line) that the entry occupies. For example, if

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 & 4 \\ 8 & 6 & 5 & -9 \\ -3 & 8 & 3 & 7 \end{pmatrix}, \quad \text{then} \quad a_{11} = 2,\ a_{12} = 1,\ \ldots,\ a_{34} = 7. \tag{1.2.6}$$

A ***submatrix*** of a given matrix $\mathbf{A}$ is an array obtained by deleting any combination of rows and columns from $\mathbf{A}$. For example, $\mathbf{B} = \begin{pmatrix} 2 & 4 \\ -3 & 7 \end{pmatrix}$ is a submatrix of the matrix $\mathbf{A}$ in (1.2.6) because $\mathbf{B}$ is the result of deleting the second row and the second and third columns of $\mathbf{A}$.

Matrix $\mathbf{A}$ is said to have ***shape*** or ***size*** $m \times n$—pronounced "m by n"—whenever $\mathbf{A}$ has exactly m rows and n columns. For example, the matrix in (1.2.6) is a 3×4 matrix. By agreement, 1×1 matrices are identified with scalars and vice versa. To emphasize that matrix $\mathbf{A}$ has shape $m \times n$, subscripts are sometimes placed on $\mathbf{A}$ as $\mathbf{A}_{m\times n}$. Whenever $m = n$ (i.e., when $\mathbf{A}$ has the same number of rows as columns), $\mathbf{A}$ is called a ***square matrix.*** Otherwise, $\mathbf{A}$ is said to be ***rectangular.*** Matrices consisting of a single row or a single column are often called ***row vectors*** or ***column vectors***, respectively.

The symbol $\mathbf{A}_{i*}$ is used to denote the ith row, while $\mathbf{A}_{*j}$ denotes the jth column of matrix $\mathbf{A}$. For example, if $\mathbf{A}$ is the matrix in (1.2.6), then

$$\mathbf{A}_{2*} = \begin{pmatrix} 8 & 6 & 5 & -9 \end{pmatrix} \quad \text{and} \quad \mathbf{A}_{*2} = \begin{pmatrix} 1 \\ 6 \\ 8 \end{pmatrix}.$$

For a linear system of equations

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m, \end{array}$$

Gaussian elimination can be executed on the associated augmented matrix $[\mathbf{A}|\mathbf{b}]$ by performing elementary operations to the rows of $[\mathbf{A}|\mathbf{b}]$. These row operations correspond to the three elementary operations (1.2.1), (1.2.2), and (1.2.3) used to manipulate linear systems. For an $m \times n$ matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{j*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix},$$

the three types of ***elementary row operations*** on $\mathbf{M}$ are as follows.

- Type I: Interchange rows i and j to produce $\begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{j*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}$. (1.2.7)

- Type II: Replace row i by a nonzero multiple of itself to produce

$$\begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \alpha\mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}, \quad \text{where} \quad \alpha \neq 0. \tag{1.2.8}$$

- Type III: Replace row j by a combination of itself plus a multiple of row i to produce

$$\begin{pmatrix} \mathbf{M}_{1*} \\ \vdots \\ \mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{j*} + \alpha\mathbf{M}_{i*} \\ \vdots \\ \mathbf{M}_{m*} \end{pmatrix}. \tag{1.2.9}$$

To solve the system (1.2.4) by using elementary row operations, start with the associated augmented matrix $[\mathbf{A}|\mathbf{b}]$ and triangularize the coefficient matrix $\mathbf{A}$ by performing exactly the same sequence of row operations that corresponds to the elementary operations executed on the equations themselves:

$$\left(\begin{array}{rrr|r} 2 & 1 & 1 & 1 \\ 6 & 2 & 1 & -1 \\ -2 & 2 & 1 & 7 \end{array}\right) \begin{array}{c} R_2 - 3R_1 \\ R_3 + R_1 \end{array} \longrightarrow \left(\begin{array}{rrr|r} 2 & 1 & 1 & 1 \\ 0 & -1 & -2 & -4 \\ 0 & 3 & 2 & 8 \end{array}\right) R_3 + 3R_2$$

$$\longrightarrow \left(\begin{array}{rrr|r} 2 & 1 & 1 & 1 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & -4 & -4 \end{array}\right).$$

The final array represents the triangular system

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

that is solved by back substitution as described earlier. In general, if an $n \times n$ system has been triangularized to the form

$$\left(\begin{array}{cccc|c} t_{11} & t_{12} & \cdots & t_{1n} & c_1 \\ 0 & t_{22} & \cdots & t_{2n} & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t_{nn} & c_n \end{array}\right) \tag{1.2.10}$$

in which each $t_{ii} \neq 0$ (i.e., there are no zero pivots), then the general algorithm for back substitution is as follows.

Algorithm for Back Substitution

Determine the x_i's from (1.2.10) by first setting $x_n = c_n/t_{nn}$ and then recursively computing

$$x_i = \frac{1}{t_{ii}}\left(c_i - t_{i,i+1}x_{i+1} - t_{i,i+2}x_{i+2} - \cdots - t_{in}x_n\right)$$

for $i = n-1,\, n-2, \ldots, 2,\, 1$.

One way to gauge the efficiency of an algorithm is to count the number of arithmetical operations required.[3] For a variety of reasons, no distinction is made between additions and subtractions, and no distinction is made between multiplications and divisions. Furthermore, multiplications/divisions are usually counted separately from additions/subtractions. Even if you do not work through the details, it is important that you be aware of the operational counts for Gaussian elimination with back substitution so that you will have a basis for comparison when other algorithms are encountered.

Gaussian Elimination Operation Counts

Gaussian elimination with back substitution applied to an $n \times n$ system requires

$$\frac{n^3}{3} + n^2 - \frac{n}{3} \quad \text{multiplications/divisions}$$

and

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \quad \text{additions/subtractions.}$$

As n grows, the $n^3/3$ term dominates each of these expressions. Therefore, the important thing to remember is that Gaussian elimination with back substitution on an $n \times n$ system requires about $n^3/3$ multiplications/divisions and about the same number of additions/subtractions.

[3] Operation counts alone may no longer be as important as they once were in gauging the efficiency of an algorithm. Older computers executed instructions sequentially, whereas some contemporary machines are capable of executing instructions in parallel so that different numerical tasks can be performed simultaneously. An algorithm that lends itself to parallelism may have a higher operational count but might nevertheless run faster on a parallel machine than an algorithm with a lesser operational count that cannot take advantage of parallelism.

Example 1.2.1

Problem: Solve the following system using Gaussian elimination with back substitution:

$$\begin{aligned} v - w &= 3, \\ -2u + 4v - w &= 1, \\ -2u + 5v - 4w &= -2. \end{aligned}$$

Solution: The associated augmented matrix is

$$\left(\begin{array}{rrr|r} 0 & 1 & -1 & 3 \\ -2 & 4 & -1 & 1 \\ -2 & 5 & -4 & -2 \end{array}\right).$$

Since the first pivotal position contains 0, interchange rows one and two before eliminating below the first pivot:

$$\left(\begin{array}{rrr|r} 0 & 1 & -1 & 3 \\ -2 & 4 & -1 & 1 \\ -2 & 5 & -4 & -2 \end{array}\right) \xrightarrow{\textit{Interchange } R_1 \textit{ and } R_2} \left(\begin{array}{rrr|r} -2 & 4 & -1 & 1 \\ 0 & 1 & -1 & 3 \\ -2 & 5 & -4 & -2 \end{array}\right) R_3 - R_1$$

$$\longrightarrow \left(\begin{array}{rrr|r} -2 & 4 & -1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 1 & -3 & -3 \end{array}\right) R_3 - R_2 \longrightarrow \left(\begin{array}{rrr|r} -2 & 4 & -1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & -2 & -6 \end{array}\right).$$

Back substitution yields

$$\begin{aligned} w &= \frac{-6}{-2} = 3, \\ v &= 3 + w = 3 + 3 = 6, \\ u &= \frac{1}{-2}(1 - 4v + w) = \frac{1}{-2}(1 - 24 + 3) = 10. \end{aligned}$$

Exercises for section 1.2

1.2.1. Use Gaussian elimination with back substitution to solve the following system:

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1 + 2x_2 + 2x_3 &= 1, \\ x_1 + 2x_2 + 3x_3 &= 1. \end{aligned}$$

1.2.2. Apply Gaussian elimination with back substitution to the following system:

$$\begin{aligned} 2x_1 - x_2 \phantom{{}- x_3} &= 0, \\ -x_1 + 2x_2 - x_3 &= 0, \\ -x_2 + x_3 &= 1. \end{aligned}$$

1.2.3. Use Gaussian elimination with back substitution to solve the following system:

$$\begin{aligned} 4x_2 - 3x_3 &= 3, \\ -x_1 + 7x_2 - 5x_3 &= 4, \\ -x_1 + 8x_2 - 6x_3 &= 5. \end{aligned}$$

1.2.4. Solve the following system:

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1, \\ x_1 + x_2 + 3x_3 + 3x_4 &= 3, \\ x_1 + x_2 + 2x_3 + 3x_4 &= 3, \\ x_1 + 3x_2 + 3x_3 + 3x_4 &= 4. \end{aligned}$$

1.2.5. Consider the following three systems where the coefficients are the same for each system, but the right-hand sides are different (this situation occurs frequently):

$$\begin{array}{r|c|c} 4x - 8y + 5z = 1 & 0 & 0, \\ 4x - 7y + 4z = 0 & 1 & 0, \\ 3x - 4y + 2z = 0 & 0 & 1. \end{array}$$

Solve all three systems at one time by performing Gaussian elimination on an augmented matrix of the form

$$\left[\mathbf{A} \mid \mathbf{b_1} \mid \mathbf{b_2} \mid \mathbf{b_3}\right].$$

1.2.6. Suppose that matrix $\mathbf{B}$ is obtained by performing a sequence of row operations on matrix $\mathbf{A}$. Explain why $\mathbf{A}$ can be obtained by performing row operations on $\mathbf{B}$.

1.2.7. Find angles α, β, and γ such that

$$\begin{aligned} 2\sin\alpha - \cos\beta + 3\tan\gamma &= 3, \\ 4\sin\alpha + 2\cos\beta - 2\tan\gamma &= 2, \\ 6\sin\alpha - 3\cos\beta + \tan\gamma &= 9, \end{aligned}$$

where $0 \le \alpha \le 2\pi$, $0 \le \beta \le 2\pi$, and $0 \le \gamma < \pi$.

1.2.8. The following system has no solution:

$$\begin{aligned} -x_1 + 3x_2 - 2x_3 &= 1, \\ -x_1 + 4x_2 - 3x_3 &= 0, \\ -x_1 + 5x_2 - 4x_3 &= 0. \end{aligned}$$

Attempt to solve this system using Gaussian elimination and explain what occurs to indicate that the system is impossible to solve.

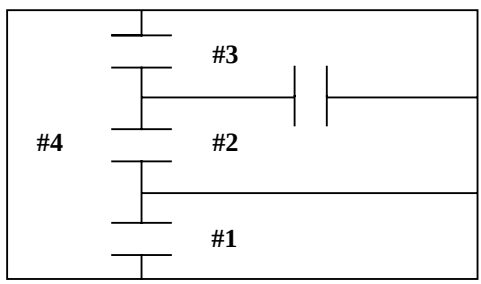
1.2.9. Attempt to solve the system

$$\begin{aligned} -x_1 + 3x_2 - 2x_3 &= 4, \\ -x_1 + 4x_2 - 3x_3 &= 5, \\ -x_1 + 5x_2 - 4x_3 &= 6, \end{aligned}$$

using Gaussian elimination and explain why this system must have infinitely many solutions.

1.2.10. By solving a 3×3 system, find the coefficients in the equation of the parabola $y = \alpha + \beta x + \gamma x^2$ that passes through the points $(1, 1)$, $(2, 2)$, and $(3, 0)$.

1.2.11. Suppose that 100 insects are distributed in an enclosure consisting of four chambers with passageways between them as shown below.



At the end of one minute, the insects have redistributed themselves. Assume that a minute is not enough time for an insect to visit more than one chamber and that at the end of a minute 40% of the insects in each chamber have not left the chamber they occupied at the beginning of the minute. The insects that leave a chamber disperse uniformly among the chambers that are directly accessible from the one they initially occupied—e.g., from #3, half move to #2 and half move to #4.

(a) If at the end of one minute there are 12, 25, 26, and 37 insects in chambers #1, #2, #3, and #4, respectively, determine what the initial distribution had to be.
(b) If the initial distribution is 20, 20, 20, 40, what is the distribution at the end of one minute?

1.2.12. Show that the three types of elementary row operations discussed on p. 8 are not independent by showing that the interchange operation (1.2.7) can be accomplished by a sequence of the other two types of row operations given in (1.2.8) and (1.2.9).

1.2.13. Suppose that $[\mathbf{A}|\mathbf{b}]$ is the augmented matrix associated with a linear system. You know that performing row operations on $[\mathbf{A}|\mathbf{b}]$ does not change the solution of the system. However, no mention of *column operations* was ever made because column operations can alter the solution.
(a) Describe the effect on the solution of a linear system when columns $\mathbf{A}_{*j}$ and $\mathbf{A}_{*k}$ are interchanged.
(b) Describe the effect when column $\mathbf{A}_{*j}$ is replaced by $\alpha\mathbf{A}_{*j}$ for $\alpha \neq 0$.
(c) Describe the effect when $\mathbf{A}_{*j}$ is replaced by $\mathbf{A}_{*j} + \alpha\mathbf{A}_{*k}$.
Hint: Experiment with a 2×2 or 3×3 system.

1.2.14. Consider the $n \times n$ ***Hilbert matrix*** defined by

$$\mathbf{H} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n-1} \end{pmatrix}.$$

Express the individual entries h_{ij} in terms of i and j.

1.2.15. Verify that the operation counts given in the text for Gaussian elimination with back substitution are correct for a general 3×3 system. If you are up to the challenge, try to verify these counts for a general $n \times n$ system.

1.2.16. Explain why a linear system can never have exactly two different solutions. Extend your argument to explain the fact that if a system has more than one solution, then it must have infinitely many different solutions.

1.3 GAUSS–JORDAN METHOD

The purpose of this section is to introduce a variation of Gaussian elimination that is known as the ***Gauss–Jordan method.*** [4] The two features that distinguish the Gauss–Jordan method from standard Gaussian elimination are as follows.

- At each step, the pivot element is forced to be 1.
- At each step, all terms *above* the pivot as well as all terms below the pivot are eliminated.

In other words, if

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right)$$

is the augmented matrix associated with a linear system, then elementary row operations are used to reduce this matrix to

$$\left(\begin{array}{cccc|c} 1 & 0 & \cdots & 0 & s_1 \\ 0 & 1 & \cdots & 0 & s_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & s_n \end{array}\right).$$

The solution then appears in the last column (i.e., $x_i = s_i$) so that this procedure circumvents the need to perform back substitution.

Example 1.3.1

Problem: Apply the Gauss–Jordan method to solve the following system:

$$\begin{aligned} 2x_1 + 2x_2 + 6x_3 &= 4, \\ 2x_1 + x_2 + 7x_3 &= 6, \\ -2x_1 - 6x_2 - 7x_3 &= -1. \end{aligned}$$

[4] Although there has been some confusion as to which Jordan should receive credit for this algorithm, it now seems clear that the method was in fact introduced by a geodesist named Wilhelm Jordan (1842–1899) and not by the more well known mathematician Marie Ennemond Camille Jordan (1838–1922), whose name is often mistakenly associated with the technique, but who is otherwise correctly credited with other important topics in matrix analysis, the "Jordan canonical form" being the most notable. Wilhelm Jordan was born in southern Germany, educated in Stuttgart, and was a professor of geodesy at the technical college in Karlsruhe. He was a prolific writer, and he introduced his elimination scheme in the 1888 publication *Handbuch der Vermessungskunde.* Interestingly, a method similar to W. Jordan's variation of Gaussian elimination seems to have been discovered and described independently by an obscure Frenchman named Clasen, who appears to have published only one scientific article, which appeared in 1888—the same year as W. Jordan's *Handbuch* appeared.

Solution: The sequence of operations is indicated in parentheses and the pivots are circled.

$$\left(\begin{array}{ccc|c} 2 & 2 & 6 & 4 \\ 2 & 1 & 7 & 6 \\ -2 & -6 & -7 & -1 \end{array}\right) R_1/2 \longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 3 & 2 \\ 2 & 1 & 7 & 6 \\ -2 & -6 & -7 & -1 \end{array}\right) \begin{array}{l} \\ R_2 - 2R_1 \\ R_3 + 2R_1 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 3 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -4 & -1 & 3 \end{array}\right) (-R_2) \longrightarrow \left(\begin{array}{ccc|c} 1 & 1 & 3 & 2 \\ 0 & 1 & -1 & -2 \\ 0 & -4 & -1 & 3 \end{array}\right) \begin{array}{l} R_1 - R_2 \\ \\ R_3 + 4R_2 \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 4 & 4 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & -5 & -5 \end{array}\right) -R_3/5 \longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 4 & 4 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 1 & 1 \end{array}\right) \begin{array}{l} R_1 - 4R_3 \\ R_2 + R_3 \\ \\ \end{array}$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{array}\right).$$

Therefore, the solution is $\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$.

On the surface it may seem that there is little difference between the Gauss–Jordan method and Gaussian elimination with back substitution because eliminating terms above the pivot with Gauss–Jordan seems equivalent to performing back substitution. But this is not correct. Gauss–Jordan requires more arithmetic than Gaussian elimination with back substitution.

Gauss–Jordan Operation Counts

For an $n \times n$ system, the Gauss–Jordan procedure requires

$$\frac{n^3}{2} + \frac{n^2}{2} \quad \text{multiplications/divisions}$$

and

$$\frac{n^3}{2} - \frac{n}{2} \quad \text{additions/subtractions.}$$

In other words, the Gauss–Jordan method requires about $n^3/2$ multiplications/divisions and about the same number of additions/subtractions.

Recall from the previous section that Gaussian elimination with back substitution requires only about $n^3/3$ multiplications/divisions and about the same

number of additions/subtractions. Compare this with the $n^3/2$ factor required by the Gauss–Jordan method, and you can see that Gauss–Jordan requires about 50% *more* effort than Gaussian elimination with back substitution. For small systems of the textbook variety (e.g., $n = 3$), these comparisons do not show a great deal of difference. However, in practical work, the systems that are encountered can be quite large, and the difference between Gauss–Jordan and Gaussian elimination with back substitution can be significant. For example, if $n = 100$, then $n^3/3$ is about 333,333, while $n^3/2$ is 500,000, which is a difference of 166,667 multiplications/divisions as well as that many additions/subtractions.

Although the Gauss–Jordan method is not recommended for solving linear systems that arise in practical applications, it does have some theoretical advantages. Furthermore, it can be a useful technique for tasks other than computing solutions to linear systems. We will make use of the Gauss–Jordan procedure when matrix inversion is discussed—this is the primary reason for introducing Gauss–Jordan.

Exercises for section 1.3

1.3.1. Use the Gauss–Jordan method to solve the following system:

$$\begin{aligned} 4x_2 - 3x_3 &= 3, \\ -x_1 + 7x_2 - 5x_3 &= 4, \\ -x_1 + 8x_2 - 6x_3 &= 5. \end{aligned}$$

1.3.2. Apply the Gauss–Jordan method to the following system:

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 1, \\ x_1 + 2x_2 + 2x_3 + 2x_4 &= 0, \\ x_1 + 2x_2 + 3x_3 + 3x_4 &= 0, \\ x_1 + 2x_2 + 3x_3 + 4x_4 &= 0. \end{aligned}$$

1.3.3. Use the Gauss–Jordan method to solve the following three systems at the same time.

$$\begin{array}{rcl|c|c} 2x_1 - x_2 & = & 1 & 0 & 0, \\ -x_1 + 2x_2 - x_3 & = & 0 & 1 & 0, \\ -x_2 + x_3 & = & 0 & 0 & 1. \end{array}$$

1.3.4. Verify that the operation counts given in the text for the Gauss–Jordan method are correct for a general 3×3 system. If you are up to the challenge, try to verify these counts for a general $n \times n$ system.

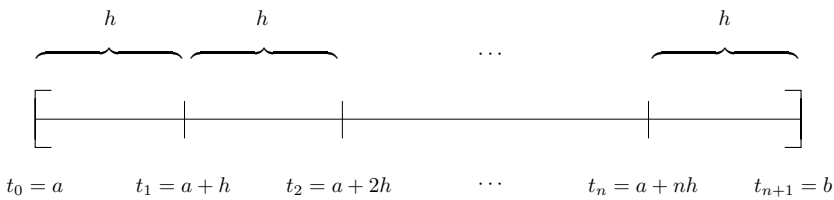
1.4 TWO-POINT BOUNDARY VALUE PROBLEMS

It was stated previously that linear systems that arise in practice can become quite large in size. The purpose of this section is to understand why this often occurs and why there is frequently a special structure to the linear systems that come from practical applications.

Given an interval $[a, b]$ and two numbers α and β, consider the general problem of trying to find a function $y(t)$ that satisfies the differential equation

$$u(t)y''(t)+v(t)y'(t)+w(t)y(t) = f(t), \quad \text{where} \quad y(a) = \alpha \text{ and } y(b) = \beta. \quad (1.4.1)$$

The functions u, v, w, and f are assumed to be known functions on $[a, b]$. Because the unknown function $y(t)$ is specified at the boundary points a and b, problem (1.4.1) is known as a ***two-point boundary value problem.*** Such problems abound in nature and are frequently very hard to handle because it is often not possible to express $y(t)$ in terms of elementary functions. Numerical methods are usually employed to approximate $y(t)$ at discrete points inside $[a, b]$. Approximations are produced by subdividing the interval $[a, b]$ into $n+1$ equal subintervals, each of length $h = (b - a)/(n + 1)$ as shown below.



Derivative approximations at the interior nodes (grid points) $t_i = a + ih$ are made by using Taylor series expansions $y(t) = \sum_{k=0}^{\infty} y^{(k)}(t_i)(t-t_i)^k/k!$ to write

$$\begin{aligned} y(t_i + h) &= y(t_i) + y'(t_i)h + \frac{y''(t_i)h^2}{2!} + \frac{y'''(t_i)h^3}{3!} + \cdots, \\ y(t_i - h) &= y(t_i) - y'(t_i)h + \frac{y''(t_i)h^2}{2!} - \frac{y'''(t_i)h^3}{3!} + \cdots, \end{aligned} \quad (1.4.2)$$

and then subtracting and adding these expressions to produce

$$y'(t_i) = \frac{y(t_i + h) - y(t_i - h)}{2h} + O(h^3)$$

and

$$y''(t_i) = \frac{y(t_i - h) - 2y(t_i) + y(t_i + h)}{h^2} + O(h^4),$$

where $O(h^p)$ denotes[5] terms containing p^{th} and higher powers of h. The

[5] Formally, a function $f(h)$ is $O(h^p)$ if $f(h)/h^p$ remains bounded as $h \to 0$, but $f(h)/h^q$ becomes unbounded if $q > p$. This means that f goes to zero as fast as h^p goes to zero.

resulting approximations

$$y'(t_i) \approx \frac{y(t_i+h) - y(t_i-h)}{2h} \text{ and } y''(t_i) \approx \frac{y(t_i-h) - 2y(t_i) + y(t_i+h)}{h^2} \tag{1.4.3}$$

are called ***centered difference approximations,*** and they are preferred over less accurate one-sided approximations such as

$$y'(t_i) \approx \frac{y(t_i+h) - y(t_i)}{h} \quad \text{or} \quad y'(t_i) \approx \frac{y(t) - y(t-h)}{h}.$$

The value $h = (b-a)/(n+1)$ is called the ***step size.*** Smaller step sizes produce better derivative approximations, so obtaining an accurate solution usually requires a small step size and a large number of grid points. By evaluating the centered difference approximations at each grid point and substituting the result into the original differential equation (1.4.1), a system of n linear equations in n unknowns is produced in which the unknowns are the values $y(t_i)$. A simple example can serve to illustrate this point.

Example 1.4.1

Suppose that $f(t)$ is a known function and consider the two-point boundary value problem

$$y''(t) = f(t) \quad \text{on } [0,1] \text{ with } y(0) = y(1) = 0.$$

The goal is to approximate the values of y at n equally spaced grid points t_i interior to $[0,1]$. The step size is therefore $h = 1/(n+1)$. For the sake of convenience, let $y_i = y(t_i)$ and $f_i = f(t_i)$. Use the approximation

$$\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \approx y''(t_i) = f_i$$

along with $y_0 = 0$ and $y_{n+1} = 0$ to produce the system of equations

$$-y_{i-1} + 2y_i - y_{i+1} \approx -h^2 f_i \quad \text{for } i = 1, 2, \ldots, n.$$

(The signs are chosen to make the 2's positive to be consistent with later developments.) The augmented matrix associated with this system is shown below:

$$\left(\begin{array}{ccccccc|c}
2 & -1 & 0 & \cdots & 0 & 0 & 0 & -h^2 f_1 \\
-1 & 2 & -1 & \cdots & 0 & 0 & 0 & -h^2 f_2 \\
0 & -1 & 2 & \cdots & 0 & 0 & 0 & -h^2 f_3 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 2 & -1 & 0 & -h^2 f_{n-2} \\
0 & 0 & 0 & \cdots & -1 & 2 & -1 & -h^2 f_{n-1} \\
0 & 0 & 0 & \cdots & 0 & -1 & 2 & -h^2 f_n
\end{array}\right).$$

By solving this system, approximate values of the unknown function y at the grid points t_i are obtained. Larger values of n produce smaller values of h and hence better approximations to the exact values of the y_i's.

Notice the pattern of the entries in the coefficient matrix in the above example. The nonzero elements occur only on the subdiagonal, main-diagonal, and superdiagonal lines—such a system (or matrix) is said to be ***tridiagonal.*** This is characteristic in the sense that when finite difference approximations are applied to the general two-point boundary value problem, a tridiagonal system is the result.

Tridiagonal systems are particularly nice in that they are inexpensive to solve. When Gaussian elimination is applied, only two multiplications/divisions are needed at each step of the triangularization process because there is at most only one nonzero entry below and to the right of each pivot. Furthermore, Gaussian elimination preserves all of the zero entries that were present in the original tridiagonal system. This makes the back substitution process cheap to execute because there are at most only two multiplications/divisions required at each substitution step. Exercise 3.10.6 contains more details.

Exercises for section 1.4

1.4.1. Divide the interval $[0,1]$ into five equal subintervals, and apply the finite difference method in order to approximate the solution of the two-point boundary value problem

$$y''(t) = 125t, \qquad y(0) = y(1) = 0$$

at the four interior grid points. Compare your approximate values at the grid points with the exact solution at the grid points. **Note:** You should not expect very accurate approximations with only four interior grid points.

1.4.2. Divide $[0,1]$ into $n+1$ equal subintervals, and apply the finite difference approximation method to derive the linear system associated with the two-point boundary value problem

$$y''(t) - y'(t) = f(t), \qquad y(0) = y(1) = 0.$$

1.4.3. Divide $[0,1]$ into five equal subintervals, and approximate the solution to

$$y''(t) - y'(t) = 125t, \qquad y(0) = y(1) = 0$$

at the four interior grid points. Compare the approximations with the exact values at the grid points.

1.5 MAKING GAUSSIAN ELIMINATION WORK

Now that you understand the basic Gaussian elimination technique, it's time to turn it into a practical algorithm that can be used for realistic applications. For pencil and paper computations where you are doing exact arithmetic, the strategy is to keep things as simple as possible (like avoiding messy fractions) in order to minimize those "stupid arithmetic errors" we are all prone to make. But very few problems in the real world are of the textbook variety, and practical applications involving linear systems usually demand the use of a computer. Computers don't care about messy fractions, and they don't introduce errors of the "stupid" variety. Computers produce a more predictable kind of error, called *roundoff error,* and it's important[6] to spend a little time up front to understand this kind of error and its effects on solving linear systems.

Numerical computation in digital computers is performed by approximating the infinite set of real numbers with a finite set of numbers as described below.

Floating-Point Numbers

A t-digit, base-β ***floating-point number*** has the form

$$f = \pm .d_1 d_2 \cdots d_t \times \beta^{\epsilon} \quad \text{with} \quad d_1 \neq 0,$$

where the base β, the exponent ϵ, and the digits $0 \leq d_i \leq \beta - 1$ are integers. For internal machine representation, $\beta = 2$ (binary representation) is standard, but for pencil-and-paper examples it's more convenient to use $\beta = 10$. The value of t, called the ***precision,*** and the exponent ϵ can vary with the choice of hardware and software.

Floating-point numbers are just adaptations of the familiar concept of scientific notation where $\beta = 10$, which will be the value used in our examples. For any fixed set of values for t, β, and ϵ, the corresponding set $\mathcal{F}$ of floating-point numbers is necessarily a finite set, so some real numbers can't be found in $\mathcal{F}$. There is more than one way of approximating real numbers with floating-point numbers. For the remainder of this text, the following common *rounding convention* is adopted. Given a real number x, the floating-point approximation $fl(x)$ is defined to be the nearest element in $\mathcal{F}$ to x, and in case of a tie we round away from 0. This means that for t-digit precision with $\beta = 10$, we need

[6] The computer has been the single most important scientific and technological development of our century and has undoubtedly altered the course of science for all future time. The prospective young scientist or engineer who passes through a contemporary course in linear algebra and matrix theory and fails to learn at least the elementary aspects of what is involved in solving a practical linear system with a computer is missing a fundamental tool of applied mathematics.

to look at digit d_{t+1} in $x = .d_1 d_2 \cdots d_t d_{t+1} \cdots \times 10^{\epsilon}$ (making sure $d_1 \neq 0$) and then set

$$fl(x) = \begin{cases} .d_1 d_2 \cdots d_t \times 10^{\epsilon} & \text{if } d_{t+1} < 5, \\ ([.d_1 d_2 \cdots d_t] + 10^{-t}) \times 10^{\epsilon} & \text{if } d_{t+1} \geq 5. \end{cases}$$

For example, in 2-digit, base-10 floating-point arithmetic,

$$fl\,(3/80) = fl(.0375) = fl(.375 \times 10^{-1}) = .38 \times 10^{-1} = .038.$$

By considering $\eta = 1/3$ and $\xi = 3$ with t-digit base-10 arithmetic, it's easy to see that

$$fl(\eta + \xi) \neq fl(\eta) + fl(\xi) \quad \text{and} \quad fl(\eta\xi) \neq fl(\eta) fl(\xi).$$

Furthermore, several familiar rules of real arithmetic do not hold for floating-point arithmetic—associativity is one outstanding example. This, among other reasons, makes the analysis of floating-point computation difficult. It also means that you must be careful when working the examples and exercises in this text because although most calculators and computers can be instructed to display varying numbers of digits, most have a fixed internal precision with which all calculations are made before numbers are displayed, and this internal precision cannot be altered. Almost certainly, the internal precision of your calculator or computer is greater than the precision called for by the examples and exercises in this text. This means that each time you perform a t-digit calculation, you should manually round the result to t *significant* digits and reenter the rounded number before proceeding to the next calculation. In other words, don't "chain" operations in your calculator or computer.

To understand how to execute Gaussian elimination using floating-point arithmetic, let's compare the use of exact arithmetic with the use of 3-digit base-10 arithmetic to solve the following system:

$$\begin{aligned} 47x + 28y &= 19, \\ 89x + 53y &= 36. \end{aligned}$$

Using Gaussian elimination with exact arithmetic, we multiply the first equation by the multiplier $m = 89/47$ and subtract the result from the second equation to produce

$$\left(\begin{array}{cc|c} 47 & 28 & 19 \\ 0 & -1/47 & 1/47 \end{array}\right).$$

Back substitution yields the *exact solution*

$$x = 1 \quad \text{and} \quad y = -1.$$

Using 3-digit arithmetic, the multiplier is

$$fl(m) = fl\left(\frac{89}{47}\right) = .189 \times 10^1 = 1.89.$$

Since

$$fl\Big(fl(m)fl(47)\Big) = fl(1.89 \times 47) = .888 \times 10^2 = 88.8,$$

$$fl\Big(fl(m)fl(28)\Big) = fl(1.89 \times 28) = .529 \times 10^2 = 52.9,$$

$$fl\Big(fl(m)fl(19)\Big) = fl(1.89 \times 19) = .359 \times 10^2 = 35.9,$$

the first step of 3-digit Gaussian elimination is as shown below:

$$\left(\begin{array}{cc|c} 47 & 28 & 19 \\ fl(89-88.8) & fl(53-52.9) & fl(36-35.9) \end{array}\right)$$

$$= \left(\begin{array}{cc|c} 47 & 28 & 19 \\ .2 & .1 & .1 \end{array}\right).$$

The goal is to triangularize the system—to produce a zero in the circled (2,1)-position—but this cannot be accomplished with 3-digit arithmetic. Unless the circled value $.2$ is replaced by 0, back substitution cannot be executed. *Henceforth, we will agree simply to enter* 0 *in the position that we are trying to annihilate,* regardless of the value of the floating-point number that might actually appear. The value of the position being annihilated is generally not even computed. For example, don't even bother computing

$$fl\Big[89 - fl\big(fl(m)fl(47)\big)\Big] = fl(89-88.8) = .2$$

in the above example. Hence the result of 3-digit Gaussian elimination for this example is

$$\left(\begin{array}{cc|c} 47 & 28 & 19 \\ 0 & .1 & .1 \end{array}\right).$$

Apply 3-digit back substitution to obtain the 3-digit floating-point solution

$$y = fl\left(\frac{.1}{.1}\right) = 1,$$

$$x = fl\left(\frac{19-28}{47}\right) = fl\left(\frac{-9}{47}\right) = -.191.$$

The vast discrepancy between the exact solution $(1,-1)$ and the 3-digit solution $(-.191, 1)$ illustrates some of the problems we can expect to encounter while trying to solve linear systems with floating-point arithmetic. Sometimes using a higher precision may help, but this is not always possible because on all machines there are natural limits that make extended precision arithmetic impractical past a certain point. Even if it is possible to increase the precision, it

may not buy you very much because there are many cases for which an increase in precision does not produce a comparable decrease in the accumulated roundoff error. Given any particular precision (say, t), it is not difficult to provide examples of linear systems for which the computed t-digit solution is just as bad as the one in our 3-digit example above.

Although the effects of rounding can almost never be eliminated, there are some simple techniques that can help to minimize these machine induced errors.

Partial Pivoting

At each step, search the positions on and below the pivotal position for the coefficient of *maximum magnitude.* If necessary perform the appropriate row interchange to bring this maximal coefficient into the pivotal position. Illustrated below is the third step in a typical case:

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{\text{S}} & * & * & * \\ 0 & 0 & S & * & * & * \\ 0 & 0 & S & * & * & * \end{array}\right).$$

Search the positions in the third column marked "S" for the coefficient of maximal magnitude and, if necessary, interchange rows to bring this coefficient into the circled pivotal position. Simply stated, the strategy is to maximize the magnitude of the pivot at each step by using only row interchanges.

On the surface, it is probably not apparent why partial pivoting should make a difference. The following example not only shows that partial pivoting can indeed make a great deal of difference, but it also indicates what makes this strategy effective.

Example 1.5.1

It is easy to verify that the exact solution to the system

$$\begin{aligned} -10^{-4}x + y &= 1, \\ x + y &= 2, \end{aligned}$$

is given by

$$x = \frac{1}{1.0001} \quad \text{and} \quad y = \frac{1.0002}{1.0001}.$$

If 3-digit arithmetic *without* partial pivoting is used, then the result is

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) R_2 + 10^4 R_1 \longrightarrow \left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 0 & 10^4 & 10^4 \end{array}\right)$$

because

$$fl(1 + 10^4) = fl(.10001 \times 10^5) = .100 \times 10^5 = 10^4 \tag{1.5.1}$$

and

$$fl(2 + 10^4) = fl(.10002 \times 10^5) = .100 \times 10^5 = 10^4. \tag{1.5.2}$$

Back substitution now produces

$$x = 0 \quad \text{and} \quad y = 1.$$

Although the computed solution for y is close to the exact solution for y, the computed solution for x is not very close to the exact solution for x—the computed solution for x is certainly not accurate to three significant figures as you might hope. If 3-digit arithmetic *with* partial pivoting is used, then the result is

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} 1 & 1 & 2 \\ -10^{-4} & 1 & 1 \end{array}\right) R_2 + 10^{-4} R_1$$
$$\longrightarrow \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array}\right)$$

because

$$fl(1 + 10^{-4}) = fl(.10001 \times 10^1) = .100 \times 10^1 = 1 \tag{1.5.3}$$

and

$$fl(1 + 2 \times 10^{-4}) = fl(.10002 \times 10^1) = .100 \times 10^1 = 1. \tag{1.5.4}$$

This time, back substitution produces the computed solution

$$x = 1 \quad \text{and} \quad y = 1,$$

which is as close to the exact solution as one can reasonably expect—the computed solution agrees with the exact solution to three significant digits.

Why did partial pivoting make a difference? The answer lies in comparing (1.5.1) and (1.5.2) with (1.5.3) and (1.5.4).

Without partial pivoting the multiplier is 10^4, and this is so large that it completely swamps the arithmetic involving the relatively smaller numbers 1 and 2 and prevents them from being taken into account. That is, the smaller numbers 1 and 2 are "blown away" as though they were never present so that our 3-digit computer produces the *exact* solution to another system, namely,

$$\left(\begin{array}{cc|c} -10^{-4} & 1 & 1 \\ 1 & 0 & 0 \end{array}\right),$$

which is quite different from the original system. *With* partial pivoting the multiplier is 10^{-4}, and this is small enough so that it does not swamp the numbers 1 and 2. In this case, the 3-digit computer produces the *exact* solution to the system $\left(\begin{array}{cc|c} 0 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right)$, which is close to the original system.[7]

In summary, the villain in Example 1.5.1 is the large multiplier that prevents some smaller numbers from being fully accounted for, thereby resulting in the exact solution of another system that is very different from the original system. By maximizing the magnitude of the pivot at each step, we minimize the magnitude of the associated multiplier thus helping to control the growth of numbers that emerge during the elimination process. This in turn helps circumvent some of the effects of roundoff error. The problem of growth in the elimination procedure is more deeply analyzed on p. 348.

When partial pivoting is used, no multiplier ever exceeds 1 in magnitude. To see that this is the case, consider the following two typical steps in an elimination procedure:

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{p} & * & * & * \\ 0 & 0 & q & * & * & * \\ 0 & 0 & r & * & * & * \end{array}\right) \begin{array}{l} \\ \\ \\ R_4 - (q/p)R_3 \\ R_5 - (r/p)R_3 \end{array} \longrightarrow \left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{p} & * & * & * \\ 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * & * \end{array}\right).$$

The pivot is p, while q/p and r/p are the multipliers. If partial pivoting has been employed, then $|p| \geq |q|$ and $|p| \geq |r|$ so that

$$\left|\frac{q}{p}\right| \leq 1 \quad \text{and} \quad \left|\frac{r}{p}\right| \leq 1.$$

By guaranteeing that no multiplier exceeds 1 in magnitude, the possibility of producing relatively large numbers that can swamp the significance of smaller numbers is much reduced, but not completely eliminated. To see that there is still more to be done, consider the following example.

Example 1.5.2

The exact solution to the system

$$\begin{aligned} -10x + 10^5 y &= 10^5, \\ x + \quad\; y &= 2, \end{aligned}$$

[7] Answering the question, "What system have I really solved (i.e., obtained the exact solution of), and how close is this system to the original system," is called *backward error analysis,* as opposed to forward analysis in which one tries to answer the question, "How close will a computed solution be to the exact solution?" Backward analysis has proven to be an effective way to analyze the numerical stability of algorithms.

is given by

$$x = \frac{1}{1.0001} \quad \text{and} \quad y = \frac{1.0002}{1.0001}.$$

Suppose that 3-digit arithmetic with partial pivoting is used. Since $|-10| > 1$, no interchange is called for and we obtain

$$\left(\begin{array}{cc|c} -10 & 10^5 & 10^5 \\ 1 & 1 & 2 \end{array}\right) R_2 + 10^{-1}R_1 \longrightarrow \left(\begin{array}{cc|c} -10 & 10^5 & 10^5 \\ 0 & 10^4 & 10^4 \end{array}\right)$$

because

$$fl(1 + 10^4) = fl(.10001 \times 10^5) = .100 \times 10^5 = 10^4$$

and

$$fl(2 + 10^4) = fl(.10002 \times 10^5) = .100 \times 10^5 = 10^4.$$

Back substitution yields

$$x = 0 \quad \text{and} \quad y = 1,$$

which must be considered to be very bad—the computed 3-digit solution for y is not too bad, but the computed 3-digit solution for x is terrible!

What is the source of difficulty in Example 1.5.2? This time, the multiplier cannot be blamed. The trouble stems from the fact that the first equation contains coefficients that are much larger than the coefficients in the second equation. That is, there is a problem of *scale* due to the fact that the coefficients are of different orders of magnitude. Therefore, we should somehow rescale the system before attempting to solve it.

If the first equation in the above example is rescaled to insure that the coefficient of maximum magnitude is a 1, which is accomplished by multiplying the first equation by 10^{-5}, then the system given in Example 1.5.1 is obtained, and we know from that example that partial pivoting produces a very good approximation to the exact solution.

This points to the fact that the success of partial pivoting can hinge on maintaining the proper scale among the coefficients. Therefore, the second refinement needed to make Gaussian elimination practical is a reasonable scaling strategy. Unfortunately, there is no known scaling procedure that will produce optimum results for every possible system, so we must settle for a strategy that will work most of the time. The strategy is to combine ***row scaling***—multiplying selected rows by nonzero multipliers—with ***column scaling***—multiplying selected columns of the coefficient matrix $\mathbf{A}$ by nonzero multipliers.

Row scaling doesn't alter the exact solution, but column scaling does—see Exercise 1.2.13(b). Column scaling is equivalent to changing the units of the kth unknown. For example, if the units of the kth unknown x_k in $[\mathbf{A}|\mathbf{b}]$ are millimeters, and if the kth column of $\mathbf{A}$ is multiplied by $.001$, then the kth unknown in the scaled system $[\hat{\mathbf{A}} \mid \mathbf{b}]$ is $\hat{x}_i = 1000x_i$, and thus the units of the scaled unknown $\hat{x}_k$ become meters.

Experience has shown that the following strategy for combining row scaling with column scaling usually works reasonably well.

Practical Scaling Strategy

1. Choose units that are natural to the problem and do not distort the relationships between the sizes of things. These natural units are usually self-evident, and further column scaling past this point is not ordinarily attempted.
2. Row scale the system $[\mathbf{A}|\mathbf{b}]$ so that the coefficient of maximum magnitude in each row of $\mathbf{A}$ is equal to 1. That is, divide each equation by the coefficient of maximum magnitude.

Partial pivoting together with the scaling strategy described above makes Gaussian elimination with back substitution an extremely effective tool. Over the course of time, this technique has proven to be reliable for solving a majority of linear systems encountered in practical work.

Although it is not extensively used, there is an extension of partial pivoting known as *complete pivoting* which, in some special cases, can be more effective than partial pivoting in helping to control the effects of roundoff error.

Complete Pivoting

If $[\mathbf{A}|\mathbf{b}]$ is the augmented matrix at the kth step of Gaussian elimination, then search the pivotal position together with every position in $\mathbf{A}$ that is below or to the right of the pivotal position for the coefficient of maximum magnitude. If necessary, perform the appropriate row and column interchanges to bring the coefficient of maximum magnitude into the pivotal position. Shown below is the third step in a typical situation:

$$\left(\begin{array}{ccccc|c} * & * & * & * & * & * \\ 0 & * & * & * & * & * \\ 0 & 0 & \textcircled{s} & S & S & * \\ 0 & 0 & S & S & S & * \\ 0 & 0 & S & S & S & * \end{array}\right)$$

Search the positions marked "S" for the coefficient of maximal magnitude. If necessary, interchange rows and columns to bring this maximal coefficient into the circled pivotal position. Recall from Exercise 1.2.13 that the effect of a column interchange in $\mathbf{A}$ is equivalent to permuting (or renaming) the associated unknowns.

You should be able to see that complete pivoting should be at least as effective as partial pivoting. Moreover, it is possible to construct specialized examples where complete pivoting is superior to partial pivoting—a famous example is presented in Exercise 1.5.7. However, one rarely encounters systems of this nature in practice. A deeper comparison between no pivoting, partial pivoting, and complete pivoting is given on p. 348.

Example 1.5.3

Problem: Use 3-digit arithmetic together with complete pivoting to solve the following system:

$$\begin{aligned} x - \quad y &= -2, \\ -9x + 10y &= 12. \end{aligned}$$

Solution: Since 10 is the coefficient of maximal magnitude that lies in the search pattern, interchange the first and second rows and then interchange the first and second columns:

$$\left(\begin{array}{cc|c} 1 & -1 & -2 \\ -9 & 10 & 12 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} -9 & 10 & 12 \\ 1 & -1 & -2 \end{array}\right)$$
$$\longrightarrow \left(\begin{array}{cc|c} 10 & -9 & 12 \\ -1 & 1 & -2 \end{array}\right) \longrightarrow \left(\begin{array}{cc|c} 10 & -9 & 12 \\ 0 & .1 & -.8 \end{array}\right).$$

The effect of the column interchange is to rename the unknowns to $\hat{x}$ and $\hat{y}$, where $\hat{x} = y$ and $\hat{y} = x$. Back substitution yields $\hat{y} = -8$ and $\hat{x} = -6$ so that

$$x = \hat{y} = -8 \quad \text{and} \quad y = \hat{x} = -6.$$

In this case, the 3-digit solution and the exact solution agree. If only partial pivoting is used, the 3-digit solution will not be as accurate. However, if scaled partial pivoting is used, the result is the same as when complete pivoting is used.

If the cost of using complete pivoting was nearly the same as the cost of using partial pivoting, we would always use complete pivoting. However, it is not difficult to show that complete pivoting approximately doubles the cost over straight Gaussian elimination, whereas partial pivoting adds only a negligible amount. Couple this with the fact that it is extremely rare to encounter a practical system where scaled partial pivoting is not adequate while complete pivoting is, and it is easy to understand why complete pivoting is seldom used in practice. Gaussian elimination with scaled partial pivoting is the preferred method for dense systems (i.e., not a lot of zeros) of moderate size.

Exercises for section 1.5

1.5.1. Consider the following system:

$$\begin{aligned} 10^{-3}x - y &= 1, \\ x + y &= 0. \end{aligned}$$

(a) Use 3-digit arithmetic with no pivoting to solve this system.
(b) Find a system that is exactly satisfied by your solution from part (a), and note how close this system is to the original system.
(c) Now use partial pivoting and 3-digit arithmetic to solve the original system.
(d) Find a system that is exactly satisfied by your solution from part (c), and note how close this system is to the original system.
(e) Use exact arithmetic to obtain the solution to the original system, and compare the exact solution with the results of parts (a) and (c).
(f) Round the exact solution to three significant digits, and compare the result with those of parts (a) and (c).

1.5.2. Consider the following system:

$$\begin{aligned} x + \quad y &= 3, \\ -10x + 10^5 y &= 10^5. \end{aligned}$$

(a) Use 4-digit arithmetic with partial pivoting and no scaling to compute a solution.
(b) Use 4-digit arithmetic with complete pivoting and no scaling to compute a solution of the original system.
(c) This time, row scale the original system first, and then apply partial pivoting with 4-digit arithmetic to compute a solution.
(d) Now determine the exact solution, and compare it with the results of parts (a), (b), and (c).

1.5.3. With no scaling, compute the 3-digit solution of

$$\begin{aligned} -3x + \ y &= -2, \\ 10x - 3y &= 7, \end{aligned}$$

without partial pivoting and with partial pivoting. Compare your results with the exact solution.

1.5.4. Consider the following system in which the coefficient matrix is the Hilbert matrix:

$$\begin{aligned} x + \frac{1}{2}y + \frac{1}{3}z &= \frac{1}{3}, \\ \frac{1}{2}x + \frac{1}{3}y + \frac{1}{4}z &= \frac{1}{3}, \\ \frac{1}{3}x + \frac{1}{4}y + \frac{1}{5}z &= \frac{1}{5}. \end{aligned}$$

(a) First convert the coefficients to 3-digit floating-point numbers, and then use 3-digit arithmetic with partial pivoting but with no scaling to compute the solution.

(b) Again use 3-digit arithmetic, but row scale the coefficients (after converting them to floating-point numbers), and then use partial pivoting to compute the solution.

(c) Proceed as in part (b), but this time row scale the coefficients *before each elimination step.*

(d) Now use exact arithmetic on the original system to determine the exact solution, and compare the result with those of parts (a), (b), and (c).

1.5.5. To see that changing units can affect a floating-point solution, consider a mining operation that extracts silica, iron, and gold from the earth. Capital (measured in dollars), operating time (in hours), and labor (in man-hours) are needed to operate the mine. To extract a pound of silica requires \$.0055, .0011 hours of operating time, and .0093 man-hours of labor. For each pound of iron extracted, \$.095, .01 operating hours, and .025 man-hours are required. For each pound of gold extracted, \$960, 112 operating hours, and 560 man-hours are required.

(a) Suppose that during 600 hours of operation, exactly \$5000 and 3000 man-hours are used. Let x, y, and z denote the number of pounds of silica, iron, and gold, respectively, that are recovered during this period. Set up the linear system whose solution will yield the values for x, y, and z.

(b) With no scaling, use 3-digit arithmetic and partial pivoting to compute a solution $(\tilde{x}, \tilde{y}, \tilde{z})$ of the system of part (a). Then approximate the exact solution (x, y, z) by using your machine's (or calculator's) full precision with partial pivoting to solve the system in part (a), and compare this with your 3-digit solution by computing the relative error defined by

$$e_r = \frac{\sqrt{(x-\tilde{x})^2 + (y-\tilde{y})^2 + (z-\tilde{z})^2}}{\sqrt{x^2+y^2+z^2}}.$$

(c) Using 3-digit arithmetic, column scale the coefficients by changing units: convert pounds of silica to tons of silica, pounds of iron to half-tons of iron, and pounds of gold to troy ounces of gold (1 lb. = 12 troy oz.).

(d) Use 3-digit arithmetic with partial pivoting to solve the column scaled system of part (c). Then approximate the exact solution by using your machine's (or calculator's) full precision with partial pivoting to solve the system in part (c), and compare this with your 3-digit solution by computing the relative error e_r as defined in part (b).

1.5.6. Consider the system given in Example 1.5.3.

(a) Use 3-digit arithmetic with partial pivoting but with no scaling to solve the system.

(b) Now use partial pivoting with scaling. Does complete pivoting provide an advantage over scaled partial pivoting in this case?

1.5.7. Consider the following well-scaled matrix:

$$\mathbf{W}_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -1 & 1 & \ddots & 0 & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \ddots & 1 & 0 & 1 \\ -1 & -1 & -1 & \cdots & -1 & 1 & 1 \\ -1 & -1 & -1 & \cdots & -1 & -1 & 1 \end{pmatrix}.$$

(a) Reduce $\mathbf{W}_n$ to an upper-triangular form using Gaussian elimination with partial pivoting, and determine the element of maximal magnitude that emerges during the elimination procedure.

(b) Now use complete pivoting and repeat part (a).

(c) Formulate a statement comparing the results of partial pivoting with those of complete pivoting for $\mathbf{W_n}$, and describe the effect this would have in determining the t-digit solution for a system whose augmented matrix is $[\mathbf{W_n} \mid \mathbf{b}]$.

1.5.8. Suppose that $\mathbf{A}$ is an $n \times n$ matrix of real numbers that has been scaled so that each entry satisfies $|a_{ij}| \leq 1$, and consider reducing $\mathbf{A}$ to triangular form using Gaussian elimination with partial pivoting. Demonstrate that after k steps of the process, no entry can have a magnitude that exceeds 2^k. **Note:** The previous exercise shows that there are cases where it is possible for some elements to actually attain the maximum magnitude of 2^k after k steps.

1.6 ILL-CONDITIONED SYSTEMS

Gaussian elimination with partial pivoting on a properly scaled system is perhaps the most fundamental algorithm in the practical use of linear algebra. However, it is not a universal algorithm nor can it be used blindly. The purpose of this section is to make the point that when solving a linear system some discretion must always be exercised because there are some systems that are so inordinately sensitive to small perturbations that *no* numerical technique can be used with confidence.

Example 1.6.1

Consider the system

$$\begin{aligned} .835x + .667y &= .168, \\ .333x + .266y &= .067, \end{aligned}$$

for which the exact solution is

$$x = 1 \quad \text{and} \quad y = -1.$$

If $b_2 = .067$ is only slightly perturbed to become $\hat{b}_2 = .066$, then the exact solution changes dramatically to become

$$\hat{x} = -666 \quad \text{and} \quad \hat{y} = 834.$$

This is an example of a system whose solution is extremely sensitive to a small perturbation. This sensitivity is intrinsic to the system itself and is not a result of any numerical procedure. Therefore, you cannot expect some "numerical trick" to remove the sensitivity. If the exact solution is sensitive to small perturbations, then any computed solution cannot be less so, regardless of the algorithm used.

Ill-Conditioned Linear Systems

A system of linear equations is said to be ***ill-conditioned*** when some small perturbation in the system can produce relatively large changes in the exact solution. Otherwise, the system is said to be ***well-conditioned.***

It is easy to visualize what causes a 2×2 system to be ill-conditioned. Geometrically, two equations in two unknowns represent two straight lines, and the point of intersection is the solution for the system. An ill-conditioned system represents two straight lines that are almost parallel.

If two straight lines are almost parallel and if one of the lines is tilted only slightly, then the point of intersection (i.e., the solution of the associated 2×2 linear system) is drastically altered.

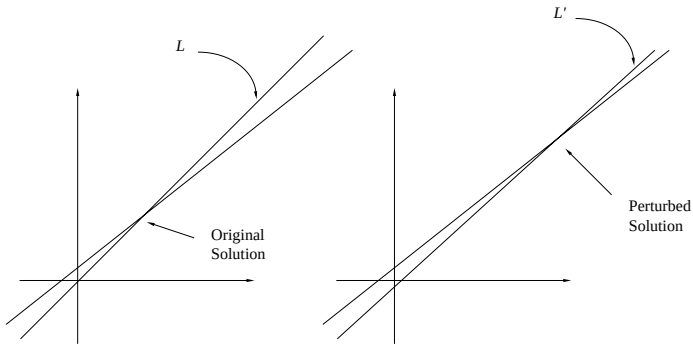


FIGURE 1.6.1

This is illustrated in Figure 1.6.1 in which line L is slightly perturbed to become line L'. Notice how this small perturbation results in a large change in the point of intersection. This was exactly the situation for the system given in Example 1.6.1. In general, ill-conditioned systems are those that represent almost parallel lines, almost parallel planes, and generalizations of these notions.

Because roundoff errors can be viewed as perturbations to the original coefficients of the system, employing even a generally good numerical technique—short of exact arithmetic—on an ill-conditioned system carries the risk of producing nonsensical results.

In dealing with an ill-conditioned system, the engineer or scientist is often confronted with a much more basic (and sometimes more disturbing) problem than that of simply trying to solve the system. Even if a minor miracle could be performed so that the exact solution could be extracted, the scientist or engineer might still have a nonsensical solution that could lead to totally incorrect conclusions. The problem stems from the fact that the coefficients are often empirically obtained and are therefore known only within certain tolerances. For an ill-conditioned system, a small uncertainty in any of the coefficients can mean an extremely large uncertainty may exist in the solution. This large uncertainty can render even the exact solution totally useless.

Example 1.6.2

Suppose that for the system

$$\begin{aligned} .835x + .667y &= b_1 \\ .333x + .266y &= b_2 \end{aligned}$$

the numbers b_1 and b_2 are the results of an experiment and must be read from the dial of a test instrument. Suppose that the dial can be read to within a

tolerance of $\pm .001$, and assume that values for b_1 and b_2 are read as $.168$ and $.067$, respectively. This produces the ill-conditioned system of Example 1.6.1, and it was seen in that example that the exact solution of the system is

$$(x, y) = (1, -1). \tag{1.6.1}$$

However, due to the small uncertainty in reading the dial, we have that

$$.167 \leq b_1 \leq .169 \quad \text{and} \quad .066 \leq b_2 \leq .068. \tag{1.6.2}$$

For example, this means that the solution associated with the reading $(b_1, b_2) = (.168, .067)$ is just as valid as the solution associated with the reading $(b_1, b_2) = (.167, .068)$, or the reading $(b_1, b_2) = (.169, .066)$, or any other reading falling in the range (1.6.2). For the reading $(b_1, b_2) = (.167, .068)$, the exact solution is

$$(x, y) = (934, -1169), \tag{1.6.3}$$

while for the other reading $(b_1, b_2) = (.169, .066)$, the exact solution is

$$(x, y) = (-932, 1167). \tag{1.6.4}$$

Would you be willing to be the first to fly in the plane or drive across the bridge whose design incorporated a solution to this problem? I wouldn't! There is just too much uncertainty. Since no one of the solutions (1.6.1), (1.6.3), or (1.6.4) can be preferred over any of the others, it is conceivable that totally different designs might be implemented depending on how the technician reads the last significant digit on the dial. Due to the ill-conditioned nature of an associated linear system, the successful design of the plane or bridge may depend on blind luck rather than on scientific principles.

Rather than trying to extract accurate solutions from ill-conditioned systems, engineers and scientists are usually better off investing their time and resources in trying to redesign the associated experiments or their data collection methods so as to avoid producing ill-conditioned systems.

There is one other discomforting aspect of ill-conditioned systems. It concerns what students refer to as "checking the answer" by substituting a computed solution back into the left-hand side of the original system of equations to see how close it comes to satisfying the system—that is, producing the right-hand side. More formally, if

$$x_c = (\, \xi_1 \quad \xi_2 \quad \cdots \quad \xi_n \,)$$

is a computed solution for a system

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n, \end{array}$$

then the numbers

$$r_i = a_{i1}\xi_1 + a_{i2}\xi_2 + \cdots + a_{in}\xi_n - b_i \qquad \text{for } i = 1, 2, \ldots, n$$

are called the ***residuals.*** Suppose that you compute a solution x_c and substitute it back to find that all the residuals are relatively small. Does this guarantee that x_c is close to the exact solution? Surprisingly, the answer is a resounding "no!" whenever the system is ill-conditioned.

Example 1.6.3

For the ill-conditioned system given in Example 1.6.1, suppose that somehow you compute a solution to be

$$\xi_1 = -666 \quad \text{and} \quad \xi_2 = 834.$$

If you attempt to "check the error" in this computed solution by substituting it back into the original system, then you find—using exact arithmetic—that the residuals are

$$r_1 = .835\xi_1 + .667\xi_2 - .168 = 0,$$
$$r_2 = .333\xi_1 + .266\xi_2 - .067 = -.001.$$

That is, the computed solution $(-666, 834)$ *exactly* satisfies the first equation and comes *very close* to satisfying the second. On the surface, this might seem to suggest that the computed solution should be very close to the exact solution. In fact a naive person could probably be seduced into believing that the computed solution is within $\pm.001$ of the exact solution. Obviously, this is nowhere close to being true since the exact solution is

$$x = 1 \quad \text{and} \quad y = -1.$$

This is always a shock to a student seeing this illustrated for the first time because it is counter to a novice's intuition. Unfortunately, many students leave school believing that they can always "check" the accuracy of their computations by simply substituting them back into the original equations—it is good to know that you're not among them.

This raises the question, *"How can I check a computed solution for accuracy?"* Fortunately, if the system is well-conditioned, then the residuals do indeed provide a more effective measure of accuracy (a rigorous proof along with more insight appears in Example 5.12.2 on p. 416). But this means that you must be able to answer some additional questions. For example, how can one tell beforehand if a given system is ill-conditioned? How can one measure the extent of ill-conditioning in a linear system?

One technique to determine the extent of ill-conditioning might be to experiment by slightly perturbing selected coefficients and observing how the solution

changes. If a radical change in the solution is observed for a small perturbation to some set of coefficients, then you have uncovered an ill-conditioned situation. If a given perturbation does not produce a large change in the solution, then nothing can be concluded—perhaps you perturbed the wrong set of coefficients.

By performing several such experiments using different sets of coefficients, a feel (but not a guarantee) for the extent of ill-conditioning can be obtained. This is expensive and not very satisfying. But before more can be said, more sophisticated tools need to be developed—the topics of sensitivity and conditioning are revisited on p. 127 and in Example 5.12.1 on p. 414.

Exercises for section 1.6

1.6.1. Consider the ill-conditioned system of Example 1.6.1:

$$\begin{aligned} .835x + .667y &= .168, \\ .333x + .266y &= .067. \end{aligned}$$

(a) Describe the outcome when you attempt to solve the system using 5-digit arithmetic with no scaling.
(b) Again using 5-digit arithmetic, first row scale the system before attempting to solve it. Describe to what extent this helps.
(c) Now use 6-digit arithmetic with no scaling. Compare the results with the exact solution.
(d) Using 6-digit arithmetic, compute the residuals for your solution of part (c), and interpret the results.
(e) For the same solution obtained in part (c), again compute the residuals, but use 7-digit arithmetic this time, and interpret the results.
(f) Formulate a concluding statement that summarizes the points made in parts (a)–(e).

1.6.2. Perturb the ill-conditioned system given in Exercise 1.6.1 above so as to form the following system:

$$\begin{aligned} .835x + .667y &= .1669995, \\ .333x + .266y &= .066601. \end{aligned}$$

(a) Determine the exact solution, and compare it with the exact solution of the system in Exercise 1.6.1.
(b) On the basis of the results of part (a), formulate a statement concerning the necessity for the solution of an ill-conditioned system to undergo a radical change for every perturbation of the original system.

1.6.3. Consider the two straight lines determined by the graphs of the following two equations:

$$\begin{aligned} .835x + .667y &= .168, \\ .333x + .266y &= .067. \end{aligned}$$

(a) Use 5-digit arithmetic to compute the slopes of each of the lines, and then use 6-digit arithmetic to do the same. In each case, sketch the graphs on a coordinate system.
(b) Show by diagram why a small perturbation in either of these lines can result in a large change in the solution.
(c) Describe in geometrical terms the situation that must exist in order for a system to be optimally well-conditioned.

1.6.4. Using geometric considerations, rank the following three systems according to their condition.

(a) $\begin{aligned} 1.001x - y &= .235, \\ x + .0001y &= .765. \end{aligned}$ (b) $\begin{aligned} 1.001x - y &= .235, \\ x + .9999y &= .765. \end{aligned}$

(c) $\begin{aligned} 1.001x + y &= .235, \\ x + .9999y &= .765. \end{aligned}$

1.6.5. Determine the exact solution of the following system:

$$\begin{aligned} 8x + 5y + 2z &= 15, \\ 21x + 19y + 16z &= 56, \\ 39x + 48y + 53z &= 140. \end{aligned}$$

Now change 15 to 14 in the first equation and again solve the system with exact arithmetic. Is the system ill-conditioned?

1.6.6. Show that the system

$$\begin{aligned} v - w - x - y - z &= 0, \\ w - x - y - z &= 0, \\ x - y - z &= 0, \\ y - z &= 0, \\ z &= 1, \end{aligned}$$

is ill-conditioned by considering the following perturbed system:

$$\begin{aligned} v - w - x - y - z &= 0, \\ -\frac{1}{15}v + w - x - y - z &= 0, \\ -\frac{1}{15}v + x - y - z &= 0, \\ -\frac{1}{15}v + y - z &= 0, \\ -\frac{1}{15}v + z &= 1. \end{aligned}$$

1.6.7. Let $f(x) = \sin \pi x$ on $[0, 1]$. The object of this problem is to determine the coefficients α_i of the cubic polynomial

$$p(x) = \sum_{i=0}^{3} \alpha_i x^i$$

that is as close to $f(x)$ as possible in the sense that

$$\begin{aligned} r &= \int_0^1 [f(x) - p(x)]^2 dx \\ &= \int_0^1 [f(x)]^2 dx - 2\sum_{i=0}^{3} \alpha_i \int_0^1 x^i f(x) dx + \int_0^1 \left(\sum_{i=0}^{3} \alpha_i x^i \right)^2 dx \end{aligned}$$

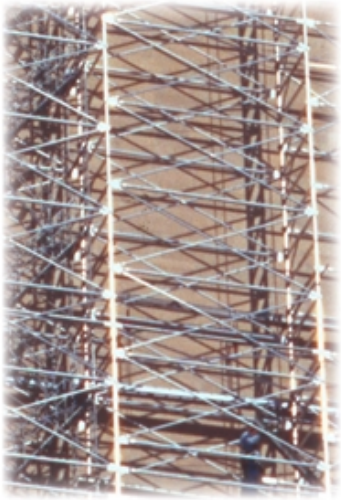
is as small as possible.

(a) In order to minimize r, impose the condition that $\partial r / \partial \alpha_i = 0$ for each $i = 0, 1, 2, 3$, and show this results in a system of linear equations whose augmented matrix is $[\mathbf{H_4} \mid \mathbf{b}]$, where $\mathbf{H_4}$ and $\mathbf{b}$ are given by

$$\mathbf{H_4} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \frac{2}{\pi} \\ \frac{1}{\pi} \\ \frac{1}{\pi} - \frac{4}{\pi^3} \\ \frac{1}{\pi} - \frac{6}{\pi^3} \end{pmatrix}.$$

Any matrix $\mathbf{H_n}$ that has the same form as $\mathbf{H_4}$ is called a ***Hilbert matrix*** of order n.

(b) Systems involving Hilbert matrices are badly ill-conditioned, and the ill-conditioning becomes worse as the size increases. Use exact arithmetic with Gaussian elimination to reduce $\mathbf{H_4}$ to triangular form. Assuming that the case in which $n = 4$ is typical, explain why a general system $[\mathbf{H_n} \mid \mathbf{b}]$ will be ill-conditioned. Notice that even complete pivoting is of no help.



CHAPTER 2

Rectangular Systems and Echelon Forms

2.1 ROW ECHELON FORM AND RANK

We are now ready to analyze more general linear systems consisting of m linear equations involving n unknowns

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m, \end{array}$$

where m may be different from n. If we do not know for sure that m and n are the same, then the system is said to be ***rectangular.*** The case $m = n$ is still allowed in the discussion—statements concerning rectangular systems also are valid for the special case of square systems.

The first goal is to extend the Gaussian elimination technique from square systems to completely general rectangular systems. Recall that for a square system with a unique solution, the pivotal positions are always located along the ***main diagonal***—the diagonal line from the upper-left-hand corner to the lower-right-hand corner—in the coefficient matrix $\mathbf{A}$ so that Gaussian elimination results in a reduction of $\mathbf{A}$ to a ***triangular matrix***, such as that illustrated below for the case $n = 4$:

$$\mathbf{T} = \begin{pmatrix} \circledast & * & * & * \\ 0 & \circledast & * & * \\ 0 & 0 & \circledast & * \\ 0 & 0 & 0 & \circledast \end{pmatrix}.$$

Remember that a pivot must always be a nonzero number. For square systems possessing a unique solution, it is a fact (proven later) that one can always bring a nonzero number into each pivotal position along the main diagonal. [8] However, in the case of a general rectangular system, it is not always possible to have the pivotal positions lying on a straight diagonal line in the coefficient matrix. This means that the final result of Gaussian elimination will *not* be triangular in form. For example, consider the following system:

$$\begin{aligned} x_1 + 2x_2 + x_3 + 3x_4 + 3x_5 &= 5, \\ 2x_1 + 4x_2 \phantom{{}+3x_3} + 4x_4 + 4x_5 &= 6, \\ x_1 + 2x_2 + 3x_3 + 5x_4 + 5x_5 &= 9, \\ 2x_1 + 4x_2 \phantom{{}+3x_3} + 4x_4 + 7x_5 &= 9. \end{aligned}$$

Focus your attention on the coefficient matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 2 & 4 & 0 & 4 & 4 \\ 1 & 2 & 3 & 5 & 5 \\ 2 & 4 & 0 & 4 & 7 \end{pmatrix}, \tag{2.1.1}$$

and ignore the right-hand side for a moment. Applying Gaussian elimination to $\mathbf{A}$ yields the following result:

$$\begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 2 & 4 & 0 & 4 & 4 \\ 1 & 2 & 3 & 5 & 5 \\ 2 & 4 & 0 & 4 & 7 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 1 & 3 & 3 \\ 0 & 0 & -2 & -2 & -2 \\ 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & -2 & -2 & 1 \end{pmatrix}.$$

In the basic elimination process, the strategy is to move down and to the right to the next pivotal position. If a zero occurs in this position, an interchange with a row below the pivotal row is executed so as to bring a nonzero number into the pivotal position. However, in this example, it is clearly impossible to bring a nonzero number into the $(2,2)$-position by interchanging the second row with a lower row.

In order to handle this situation, the elimination process is modified as follows.

[8] This discussion is for exact arithmetic. If floating-point arithmetic is used, this may no longer be true. Part (a) of Exercise 1.6.1 is one such example.

Modified Gaussian Elimination

Suppose that $\mathbf{U}$ is the augmented matrix associated with the system after $i-1$ elimination steps have been completed. To execute the ith step, proceed as follows:

- Moving from left to right in $\mathbf{U}$, locate the first column that contains a nonzero entry on or below the ith position—say it is $\mathbf{U}_{*j}$.
- The pivotal position for the ith step is the (i,j)-position.
- If necessary, interchange the ith row with a lower row to bring a nonzero number into the (i,j)-position, and then annihilate all entries below this pivot.
- If row $\mathbf{U}_{i*}$ as well as all rows in $\mathbf{U}$ below $\mathbf{U}_{i*}$ consist entirely of zeros, then the elimination process is completed.

Illustrated below is the result of applying this modified version of Gaussian elimination to the matrix given in (2.1.1).

Example 2.1.1

Problem: Apply modified Gaussian elimination to the following matrix and circle the pivot positions:

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

Solution:

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

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

Notice that the final result of applying Gaussian elimination in the above example is not a purely triangular form but rather a jagged or "stair-step" type of triangular form. Hereafter, a matrix that exhibits this stair-step structure will be said to be in ***row echelon form.***

Row Echelon Form

An $m \times n$ matrix $\mathbf{E}$ with rows $\mathbf{E}_{i*}$ and columns $\mathbf{E}_{*j}$ is said to be in ***row echelon form*** provided the following two conditions hold.

- If $\mathbf{E}_{i*}$ consists entirely of zeros, then all rows below $\mathbf{E}_{i*}$ are also entirely zero; i.e., all zero rows are at the bottom.
- If the first nonzero entry in $\mathbf{E}_{i*}$ lies in the jth position, then all entries below the ith position in columns $\mathbf{E}_{*1}, \mathbf{E}_{*2}, \ldots, \mathbf{E}_{*j}$ are zero.

These two conditions say that the nonzero entries in an echelon form must lie on or above a stair-step line that emanates from the upper-left-hand corner and slopes down and to the right. The pivots are the first nonzero entries in each row. A typical structure for a matrix in row echelon form is illustrated below with the pivots circled.

$$\begin{pmatrix} \textcircled{*} & * & * & * & * & * & * & * \\ 0 & 0 & \textcircled{*} & * & * & * & * & * \\ 0 & 0 & 0 & \textcircled{*} & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & \textcircled{*} & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Because of the flexibility in choosing row operations to reduce a matrix $\mathbf{A}$ to a row echelon form $\mathbf{E}$, the entries in $\mathbf{E}$ are not uniquely determined by $\mathbf{A}$. Nevertheless, it can be proven that the "form" of $\mathbf{E}$ is unique in the sense that *the positions of the pivots in* $\mathbf{E}$ (*and* $\mathbf{A}$) *are uniquely determined by the entries in* $\mathbf{A}$.[9] Because the pivotal positions are unique, it follows that the number of pivots, which is the same as the number of nonzero rows in $\mathbf{E}$, is also uniquely determined by the entries in $\mathbf{A}$. This number is called the ***rank***[10] of $\mathbf{A}$, and it

[9] The fact that the pivotal positions are unique should be intuitively evident. If it isn't, take the matrix given in (2.1.1) and try to force some different pivotal positions by a different sequence of row operations.

[10] The word "rank" was introduced in 1879 by the German mathematician Ferdinand Georg Frobenius (p. 662), who thought of it as the size of the largest nonzero minor determinant in $\mathbf{A}$. But the concept had been used as early as 1851 by the English mathematician James J. Sylvester (1814–1897).

is extremely important in the development of our subject.

Rank of a Matrix

Suppose $\mathbf{A}_{m\times n}$ is reduced by row operations to an echelon form $\mathbf{E}$. The ***rank*** of $\mathbf{A}$ is defined to be the number

$$\begin{aligned} rank\,(\mathbf{A}) &= \text{number of pivots} \\ &= \text{number of nonzero rows in } \mathbf{E} \\ &= \text{number of basic columns in } \mathbf{A}, \end{aligned}$$

where the ***basic columns*** of $\mathbf{A}$ are defined to be those columns in $\mathbf{A}$ that contain the pivotal positions.

Example 2.1.2

Problem: Determine the rank, and identify the basic columns in

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

Solution: Reduce $\mathbf{A}$ to row echelon form as shown below:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 3 & 6 & 3 & 4 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E}.$$

Consequently, $rank\,(\mathbf{A}) = 2$. The pivotal positions lie in the first and fourth columns so that the basic columns of $\mathbf{A}$ are $\mathbf{A}_{*1}$ and $\mathbf{A}_{*4}$. That is,

$$\text{Basic Columns} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} \right\}.$$

Pay particular attention to the fact that the basic columns are extracted from $\mathbf{A}$ and not from the row echelon form $\mathbf{E}$.

Exercises for section 2.1

2.1.1. Reduce each of the following matrices to row echelon form, determine the rank, and identify the basic columns.

$$\text{(a)}\begin{pmatrix}1&2&3&3\\2&4&6&9\\2&6&7&6\end{pmatrix}\quad\text{(b)}\begin{pmatrix}1&2&3\\2&6&8\\2&6&0\\1&2&5\\3&8&6\end{pmatrix}\quad\text{(c)}\begin{pmatrix}2&1&1&3&0&4&1\\4&2&4&4&1&5&5\\2&1&3&1&0&4&3\\6&3&4&8&1&9&5\\0&0&3&-3&0&0&3\\8&4&2&14&1&13&3\end{pmatrix}$$

2.1.2. Determine which of the following matrices are in row echelon form:

$$\text{(a)}\begin{pmatrix}1&2&3\\0&0&4\\0&1&0\end{pmatrix}.\qquad\text{(b)}\begin{pmatrix}0&0&0&0\\0&1&0&0\\0&0&0&1\end{pmatrix}.$$

$$\text{(c)}\begin{pmatrix}2&2&3&-4\\0&0&7&-8\\0&0&0&-1\end{pmatrix}.\qquad\text{(d)}\begin{pmatrix}1&2&0&0&1&0\\0&0&0&1&0&0\\0&0&0&0&0&1\\0&0&0&0&0&0\end{pmatrix}.$$

2.1.3. Suppose that $\mathbf{A}$ is an $m \times n$ matrix. Give a short explanation of why each of the following statements is true.

(a) $rank\,(\mathbf{A}) \leq \min\{m, n\}$.
(b) $rank\,(\mathbf{A}) < m$ if one row in $\mathbf{A}$ is entirely zero.
(c) $rank\,(\mathbf{A}) < m$ if one row in $\mathbf{A}$ is a multiple of another row.
(d) $rank\,(\mathbf{A}) < m$ if one row in $\mathbf{A}$ is a combination of other rows.
(e) $rank\,(\mathbf{A}) < n$ if one column in $\mathbf{A}$ is entirely zero.

2.1.4. Let $\mathbf{A} = \begin{pmatrix}.1&.2&.3\\.4&.5&.6\\.7&.8&.901\end{pmatrix}$.

(a) Use exact arithmetic to determine $rank\,(\mathbf{A})$.
(b) Now use 3-digit floating-point arithmetic (without partial pivoting or scaling) to determine $rank\,(\mathbf{A})$. This number might be called the "3-digit numerical rank."
(c) What happens if partial pivoting is incorporated?

2.1.5. How many different "forms" are possible for a 3×4 matrix that is in row echelon form?

2.1.6. Suppose that $[\mathbf{A}|\mathbf{b}]$ is reduced to a matrix $[\mathbf{E}|\mathbf{c}]$.

(a) Is $[\mathbf{E}|\mathbf{c}]$ in row echelon form if $\mathbf{E}$ is?
(b) If $[\mathbf{E}|\mathbf{c}]$ is in row echelon form, must $\mathbf{E}$ be in row echelon form?

2.2 REDUCED ROW ECHELON FORM

At each step of the Gauss–Jordan method, the pivot is forced to be a 1, and then all entries above and below the pivotal 1 are annihilated. If $\mathbf{A}$ is the coefficient matrix for a square system with a unique solution, then the end result of applying the Gauss–Jordan method to $\mathbf{A}$ is a matrix with 1's on the main diagonal and 0's everywhere else. That is,

$$\mathbf{A} \xrightarrow{\text{Gauss–Jordan}} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

But if the Gauss–Jordan technique is applied to a more general $m \times n$ matrix, then the final result is not necessarily the same as described above. The following example illustrates what typically happens in the rectangular case.

Example 2.2.1

Problem: Apply Gauss–Jordan elimination to the following 4×5 matrix and circle the pivot positions. This is the same matrix used in Example 2.1.1:

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

Solution:

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

$$\to \begin{pmatrix} 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\to \begin{pmatrix} 1 & 2 & 0 & 2 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Compare the results of this example with the results of Example 2.1.1, and notice that the "form" of the final matrix is the same in both examples, which indeed must be the case because of the uniqueness of "form" mentioned in the previous section. The only difference is in the numerical value of some of the entries. By the nature of Gauss–Jordan elimination, each pivot is 1 and all entries above and below each pivot are 0. Consequently, the row echelon form produced by the Gauss–Jordan method contains a reduced number of nonzero entries, so it seems only natural to refer to this as a ***reduced row echelon form.***[11]

Reduced Row Echelon Form

A matrix $\mathbf{E}_{m\times n}$ is said to be in ***reduced row echelon form*** provided that the following three conditions hold.

- $\mathbf{E}$ is in row echelon form.
- The first nonzero entry in each row (i.e., each pivot) is 1.
- All entries above each pivot are 0.

A typical structure for a matrix in reduced row echelon form is illustrated below, where entries marked * can be either zero or nonzero numbers:

$$\begin{pmatrix} 1 & * & 0 & 0 & * & * & 0 & * \\ 0 & 0 & 1 & 0 & * & * & 0 & * \\ 0 & 0 & 0 & 1 & * & * & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

As previously stated, if matrix $\mathbf{A}$ is transformed to a row echelon form by row operations, then the "form" is uniquely determined by $\mathbf{A}$, but the individual entries in the form are not unique. However, if $\mathbf{A}$ is transformed by row operations to a *reduced* row echelon form $\mathbf{E_A}$, then it can be shown[12] that both the "form" as well as the individual entries in $\mathbf{E_A}$ are uniquely determined by $\mathbf{A}$. In other words, the reduced row echelon form $\mathbf{E_A}$ produced from $\mathbf{A}$ is independent of whatever elimination scheme is used. Producing an unreduced form is computationally more efficient, but the uniqueness of $\mathbf{E_A}$ makes it more useful for theoretical purposes.

---

[11] In some of the older books this is called the *Hermite normal form* in honor of the French mathematician Charles Hermite (1822–1901), who, around 1851, investigated reducing matrices by row operations.

[12] A formal uniqueness proof must wait until Example 3.9.2, but you can make this intuitively clear right now with some experiments. Try to produce two different reduced row echelon forms from the same matrix.

$\mathbf{E_A}$ Notation

For a matrix $\mathbf{A}$, the symbol $\mathbf{E_A}$ will hereafter denote the unique reduced row echelon form derived from $\mathbf{A}$ by means of row operations.

Example 2.2.2

Problem: Determine $\mathbf{E_A}$, deduce $rank\,(\mathbf{A})$, and identify the basic columns of

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

Solution:

$$\begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 2 & 4 & 4 & 6 & 2 \\ 3 & 6 & 6 & 9 & 6 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 2 & 2 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 2 & 3 & 1 \\ 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

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

$$\longrightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Therefore, $rank\,(\mathbf{A}) = 3$, and $\{\mathbf{A}_{*1}, \mathbf{A}_{*3}, \mathbf{A}_{*5}\}$ are the three basic columns.

The above example illustrates another important feature of $\mathbf{E_A}$ and explains why the basic columns are indeed "basic." Each nonbasic column is expressible as a combination of basic columns. In Example 2.2.2,

$$\mathbf{A}_{*2} = 2\mathbf{A}_{*1} \quad \text{and} \quad \mathbf{A}_{*4} = \mathbf{A}_{*1} + \mathbf{A}_{*3}. \tag{2.2.1}$$

Notice that exactly the same set of relationships hold in $\mathbf{E_A}$. That is,

$$\mathbf{E}_{*2} = 2\mathbf{E}_{*1} \quad \text{and} \quad \mathbf{E}_{*4} = \mathbf{E}_{*1} + \mathbf{E}_{*3}. \tag{2.2.2}$$

This is no coincidence—it's characteristic of what happens in general. There's more to observe. The relationships between the nonbasic and basic columns in a

general matrix $\mathbf{A}$ are usually obscure, but the relationships among the columns in $\mathbf{E_A}$ are absolutely transparent. For example, notice that the multipliers used in the relationships (2.2.1) and (2.2.2) appear explicitly in the two nonbasic columns in $\mathbf{E_A}$—they are just the nonzero entries in these nonbasic columns. This is important because it means that $\mathbf{E_A}$ can be used as a "map" or "key" to discover or unlock the hidden relationships among the columns of $\mathbf{A}$.

Finally, observe from Example 2.2.2 that only the basic columns *to the left* of a given nonbasic column are needed in order to express the nonbasic column as a combination of basic columns—e.g., representing $\mathbf{A}_{*2}$ requires only $\mathbf{A}_{*1}$ and not $\mathbf{A}_{*3}$ or $\mathbf{A}_{*5}$, while representing $\mathbf{A}_{*4}$ requires only $\mathbf{A}_{*1}$ and $\mathbf{A}_{*3}$. This too is typical. For the time being, we accept the following statements to be true. A rigorous proof is given later on p. 136.

Column Relationships in A and $\mathbf{E_A}$

- Each nonbasic column $\mathbf{E}_{*k}$ in $\mathbf{E_A}$ is a combination (a sum of multiples) of the basic columns in $\mathbf{E_A}$ to the left of $\mathbf{E}_{*k}$. That is,

$$\begin{aligned}\mathbf{E}_{*k} &= \mu_1\mathbf{E}_{*b_1} + \mu_2\mathbf{E}_{*b_2} + \cdots + \mu_j\mathbf{E}_{*b_j}\\ &= \mu_1\begin{pmatrix}1\\0\\\vdots\\0\\\vdots\\0\end{pmatrix} + \mu_2\begin{pmatrix}0\\1\\\vdots\\0\\\vdots\\0\end{pmatrix} + \cdots + \mu_j\begin{pmatrix}0\\0\\\vdots\\1\\\vdots\\0\end{pmatrix} = \begin{pmatrix}\mu_1\\\mu_2\\\vdots\\\mu_j\\\vdots\\0\end{pmatrix},\end{aligned}$$

  where the $\mathbf{E}_{*b_i}$'s are the basic columns to the left of $\mathbf{E}_{*k}$ and where the multipliers μ_i are the first j entries in $\mathbf{E}_{*k}$.

- The relationships that exist among the columns of $\mathbf{A}$ are exactly the same as the relationships that exist among the columns of $\mathbf{E_A}$. In particular, if $\mathbf{A}_{*k}$ is a nonbasic column in $\mathbf{A}$, then

$$\mathbf{A}_{*k} = \mu_1\mathbf{A}_{*b_1} + \mu_2\mathbf{A}_{*b_2} + \cdots + \mu_j\mathbf{A}_{*b_j}, \tag{2.2.3}$$

  where the $\mathbf{A}_{*b_i}$'s are the basic columns to the left of $\mathbf{A}_{*k}$, and where the multipliers μ_i are as described above—the first j entries in $\mathbf{E}_{*k}$.

Example 2.2.3

Problem: Write each nonbasic column as a combination of basic columns in

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

Solution: Transform $\mathbf{A}$ to $\mathbf{E_A}$ as shown below.

$$\begin{pmatrix} 2 & -4 & -8 & 6 & 3 \\ 0 & 1 & 3 & 2 & 3 \\ 3 & -2 & 0 & 0 & 8 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & -4 & 3 & \frac{3}{2} \\ 0 & 1 & 3 & 2 & 3 \\ 3 & -2 & 0 & 0 & 8 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & -4 & 3 & \frac{3}{2} \\ 0 & 1 & 3 & 2 & 3 \\ 0 & 4 & 12 & -9 & \frac{7}{2} \end{pmatrix} \to$$

$$\begin{pmatrix} 1 & 0 & 2 & 7 & \frac{15}{2} \\ 0 & 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & -17 & -\frac{17}{2} \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 2 & 7 & \frac{15}{2} \\ 0 & 1 & 3 & 2 & 3 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 2 & 0 & 4 \\ 0 & 1 & 3 & 0 & 2 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \end{pmatrix}$$

The third and fifth columns are nonbasic. Looking at the columns in $\mathbf{E_A}$ reveals

$$\mathbf{E}_{*3} = 2\mathbf{E}_{*1} + 3\mathbf{E}_{*2} \quad \text{and} \quad \mathbf{E}_{*5} = 4\mathbf{E}_{*1} + 2\mathbf{E}_{*2} + \frac{1}{2}\mathbf{E}_{*4}.$$

The relationships that exist among the columns of $\mathbf{A}$ must be exactly the same as those in $\mathbf{E_A}$, so

$$\mathbf{A}_{*3} = 2\mathbf{A}_{*1} + 3\mathbf{A}_{*2} \quad \text{and} \quad \mathbf{A}_{*5} = 4\mathbf{A}_{*1} + 2\mathbf{A}_{*2} + \frac{1}{2}\mathbf{A}_{*4}.$$

You can easily check the validity of these equations by direct calculation.

In summary, the utility of $\mathbf{E_A}$ lies in its ability to reveal dependencies in data stored as columns in an array $\mathbf{A}$. The nonbasic columns in $\mathbf{A}$ represent redundant information in the sense that this information can always be expressed in terms of the data contained in the basic columns.

Although data compression is not the primary reason for introducing $\mathbf{E_A}$, the application to these problems is clear. For a large array of data, it may be more efficient to store only "independent data" (i.e., the basic columns of $\mathbf{A}$) along with the nonzero multipliers μ_i obtained from the nonbasic columns in $\mathbf{E_A}$. Then the redundant data contained in the nonbasic columns of $\mathbf{A}$ can always be reconstructed if and when it is called for.

Exercises for section 2.2

2.2.1. Determine the reduced row echelon form for each of the following matrices and then express each nonbasic column in terms of the basic columns:

$$\text{(a)} \quad \begin{pmatrix} 1 & 2 & 3 & 3 \\ 2 & 4 & 6 & 9 \\ 2 & 6 & 7 & 6 \end{pmatrix}, \qquad \text{(b)} \quad \begin{pmatrix} 2 & 1 & 1 & 3 & 0 & 4 & 1 \\ 4 & 2 & 4 & 4 & 1 & 5 & 5 \\ 2 & 1 & 3 & 1 & 0 & 4 & 3 \\ 6 & 3 & 4 & 8 & 1 & 9 & 5 \\ 0 & 0 & 3 & -3 & 0 & 0 & 3 \\ 8 & 4 & 2 & 14 & 1 & 13 & 3 \end{pmatrix}.$$

2.2.2. Construct a matrix $\mathbf{A}$ whose reduced row echelon form is

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

Is $\mathbf{A}$ unique?

2.2.3. Suppose that $\mathbf{A}$ is an $m \times n$ matrix. Give a short explanation of why $rank\,(\mathbf{A}) < n$ whenever one column in $\mathbf{A}$ is a combination of other columns in $\mathbf{A}$.

2.2.4. Consider the following matrix:

$$\mathbf{A} = \begin{pmatrix} .1 & .2 & .3 \\ .4 & .5 & .6 \\ .7 & .8 & .901 \end{pmatrix}.$$

(a) Use exact arithmetic to determine $\mathbf{E_A}$.
(b) Now use 3-digit floating-point arithmetic (without partial pivoting or scaling) to determine $\mathbf{E_A}$ and formulate a statement concerning "near relationships" between the columns of $\mathbf{A}$.

2.2.5. Consider the matrix

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

You already know that $\mathbf{E}_{*3}$ can be expressed in terms of $\mathbf{E}_{*1}$ and $\mathbf{E}_{*2}$. However, this is not the only way to represent the column dependencies in $\mathbf{E}$. Show how to write $\mathbf{E}_{*1}$ in terms of $\mathbf{E}_{*2}$ and $\mathbf{E}_{*3}$ and then express $\mathbf{E}_{*2}$ as a combination of $\mathbf{E}_{*1}$ and $\mathbf{E}_{*3}$. **Note:** This exercise illustrates that the set of pivotal columns is not the only set that can play the role of "basic columns." Taking the basic columns to be the ones containing the pivots is a matter of convenience because everything becomes automatic that way.

2.3 CONSISTENCY OF LINEAR SYSTEMS

A system of m linear equations in n unknowns is said to be a ***consistent*** system if it possesses at least one solution. If there are no solutions, then the system is called ***inconsistent.*** The purpose of this section is to determine conditions under which a given system will be consistent.

Stating conditions for consistency of systems involving only two or three unknowns is easy. A linear equation in two unknowns represents a line in 2-space, and a linear equation in three unknowns is a plane in 3-space. Consequently, a linear system of m equations in two unknowns is consistent if and only if the m lines defined by the m equations have at least one common point of intersection. Similarly, a system of m equations in three unknowns is consistent if and only if the associated m planes have at least one common point of intersection. However, when m is large, these geometric conditions may not be easy to verify visually, and when $n > 3$, the generalizations of intersecting lines or planes are impossible to visualize with the eye.

Rather than depending on geometry to establish consistency, we use Gaussian elimination. If the associated augmented matrix $[\mathbf{A}|\mathbf{b}]$ is reduced by row operations to a matrix $[\mathbf{E}|\mathbf{c}]$ that is in row echelon form, then consistency—or lack of it—becomes evident. Suppose that somewhere in the process of reducing $[\mathbf{A}|\mathbf{b}]$ to $[\mathbf{E}|\mathbf{c}]$ a situation arises in which the only nonzero entry in a row appears on the right-hand side, as illustrated below:

$$\text{Row } i \longrightarrow \left(\begin{array}{cccccc|c} * & * & * & * & * & * & * \\ 0 & 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{array}\right) \longleftarrow \alpha \neq 0.$$

If this occurs in the ith row, then the ith equation of the associated system is

$$0x_1 + 0x_2 + \cdots + 0x_n = \alpha.$$

For $\alpha \neq 0$, this equation has no solution, and hence the original system must also be inconsistent (because row operations don't alter the solution set). The converse also holds. That is, if a system is inconsistent, then somewhere in the elimination process a row of the form

$$(\,0 \quad 0 \quad \cdots \quad 0 \quad | \quad \alpha\,), \quad \alpha \neq 0 \tag{2.3.1}$$

must appear. Otherwise, the back substitution process can be completed and a solution is produced. There is *no* inconsistency indicated when a row of the form $(0\ 0\ \cdots\ 0\,|\,0)$ is encountered. This simply says that $0 = 0$, and although

this is no help in determining the value of any unknown, it is nevertheless a true statement, so it doesn't indicate inconsistency in the system.

There are some other ways to characterize the consistency (or inconsistency) of a system. One of these is to observe that if the last column $\mathbf{b}$ in the augmented matrix $[\mathbf{A}|\mathbf{b}]$ is a nonbasic column, then no pivot can exist in the last column, and hence the system is consistent because the situation (2.3.1) cannot occur. Conversely, if the system is consistent, then the situation (2.3.1) never occurs during Gaussian elimination and consequently the last column cannot be basic. In other words, $[\mathbf{A}|\mathbf{b}]$ is consistent if and only if $\mathbf{b}$ is a nonbasic column.

Saying that $\mathbf{b}$ is a nonbasic column in $[\mathbf{A}|\mathbf{b}]$ is equivalent to saying that all basic columns in $[\mathbf{A}|\mathbf{b}]$ lie in the coefficient matrix $\mathbf{A}$. Since the number of basic columns in a matrix is the rank, consistency may also be characterized by stating that a system is consistent if and only if $rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A})$.

Recall from the previous section the fact that each nonbasic column in $[\mathbf{A}|\mathbf{b}]$ must be expressible in terms of the basic columns. Because a consistent system is characterized by the fact that the right-hand side $\mathbf{b}$ is a nonbasic column, it follows that a system is consistent if and only if the right-hand side $\mathbf{b}$ is a combination of columns from the coefficient matrix $\mathbf{A}$.

Each of the equivalent[13] ways of saying that a system is consistent is summarized below.

Consistency

Each of the following is equivalent to saying that $[\mathbf{A}|\mathbf{b}]$ is consistent.

- In row reducing $[\mathbf{A}|\mathbf{b}]$, a row of the following form never appears:

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & | & \alpha \end{pmatrix}, \quad \text{where} \quad \alpha \neq 0. \tag{2.3.2}$$

- $\mathbf{b}$ is a nonbasic column in $[\mathbf{A}|\mathbf{b}]$. (2.3.3)
- $rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A})$. (2.3.4)
- $\mathbf{b}$ is a combination of the basic columns in $\mathbf{A}$. (2.3.5)

Example 2.3.1

Problem: Determine if the following system is consistent:

$$\begin{aligned} x_1 + x_2 + 2x_3 + 2x_4 + x_5 &= 1, \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 3x_5 &= 1, \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 2x_5 &= 2, \\ 3x_1 + 5x_2 + 8x_3 + 6x_4 + 5x_5 &= 3. \end{aligned}$$

[13] Statements P and Q are said to be equivalent when (P implies Q) as well as its converse (Q implies P) are true statements. This is also the meaning of the phrase "P if and only if Q."

Solution: Apply Gaussian elimination to the augmented matrix $[\mathbf{A}|\mathbf{b}]$ as shown:

$$\left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1 \\ 2 & 2 & 4 & 4 & 3 & 1 \\ 2 & 2 & 4 & 4 & 2 & 2 \\ 3 & 5 & 8 & 6 & 5 & 3 \end{array}\right) \longrightarrow \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 2 & 0 & 2 & 0 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1 \\ 0 & 2 & 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Because a row of the form $(0 \quad 0 \quad \cdots \quad 0 \quad | \quad \alpha)$ with $\alpha \neq 0$ never emerges, the system is consistent. We might also observe that $\mathbf{b}$ is a nonbasic column in $[\mathbf{A}|\mathbf{b}]$ so that $rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A})$. Finally, by completely reducing $\mathbf{A}$ to $\mathbf{E_A}$, it is possible to verify that $\mathbf{b}$ is indeed a combination of the basic columns $\{\mathbf{A}_{*1}, \mathbf{A}_{*2}, \mathbf{A}_{*5}\}$.

Exercises for section 2.3

2.3.1. Determine which of the following systems are consistent.

(a) $\begin{aligned} x + 2y + z &= 2, \\ 2x + 4y \qquad &= 2, \\ 3x + 6y + z &= 4. \end{aligned}$ (b) $\begin{aligned} 2x + 2y + 4z &= 0, \\ 3x + 2y + 5z &= 0, \\ 4x + 2y + 6z &= 0. \end{aligned}$

(c) $\begin{aligned} x - y + z &= 1, \\ x - y - z &= 2, \\ x + y - z &= 3, \\ x + y + z &= 4. \end{aligned}$ (d) $\begin{aligned} x - y + z &= 1, \\ x - y - z &= 2, \\ x + y - z &= 3, \\ x + y + z &= 2. \end{aligned}$

(e) $\begin{aligned} 2w + x + 3y + 5z &= 1, \\ 4w + \qquad 4y + 8z &= 0, \\ w + x + 2y + 3z &= 0, \\ x + \ y + \ z &= 0. \end{aligned}$ (f) $\begin{aligned} 2w + x + 3y + 5z &= 7, \\ 4w + \qquad 4y + 8z &= 8, \\ w + x + 2y + 3z &= 5, \\ x + \ y + \ z &= 3. \end{aligned}$

2.3.2. Construct a 3×4 matrix $\mathbf{A}$ and 3×1 columns $\mathbf{b}$ and $\mathbf{c}$ such that $[\mathbf{A}|\mathbf{b}]$ is the augmented matrix for an inconsistent system, but $[\mathbf{A}|\mathbf{c}]$ is the augmented matrix for a consistent system.

2.3.3. If $\mathbf{A}$ is an $m \times n$ matrix with $rank\,(\mathbf{A}) = m$, explain why the system $[\mathbf{A}|\mathbf{b}]$ must be consistent for every right-hand side $\mathbf{b}$.

2.3.4. Consider two consistent systems whose augmented matrices are of the form $[\mathbf{A}|\mathbf{b}]$ and $[\mathbf{A}|\mathbf{c}]$. That is, they differ only on the right-hand side. Is the system associated with $[\mathbf{A} \mid \mathbf{b}+\mathbf{c}]$ also consistent? Explain why.

2.3.5. Is it possible for a parabola whose equation has the form $y = \alpha+\beta x+\gamma x^2$ to pass through the four points $(0,1)$, $(1,3)$, $(2,15)$, and $(3,37)$? Why?

2.3.6. Consider using floating-point arithmetic (without scaling) to solve the following system:

$$\begin{aligned} .835x + .667y &= .168, \\ .333x + .266y &= .067. \end{aligned}$$

(a) Is the system consistent when 5-digit arithmetic is used?
(b) What happens when 6-digit arithmetic is used?

2.3.7. In order to grow a certain crop, it is recommended that each square foot of ground be treated with 10 units of phosphorous, 9 units of potassium, and 19 units of nitrogen. Suppose that there are three brands of fertilizer on the market— say brand $\mathcal{X}$, brand $\mathcal{Y}$, and brand $\mathcal{Z}$. One pound of brand $\mathcal{X}$ contains 2 units of phosphorous, 3 units of potassium, and 5 units of nitrogen. One pound of brand $\mathcal{Y}$ contains 1 unit of phosphorous, 3 units of potassium, and 4 units of nitrogen. One pound of brand $\mathcal{Z}$ contains only 1 unit of phosphorous and 1 unit of nitrogen. Determine whether or not it is possible to meet exactly the recommendation by applying some combination of the three brands of fertilizer.

2.3.8. Suppose that an augmented matrix $[\mathbf{A}|\mathbf{b}]$ is reduced by means of Gaussian elimination to a row echelon form $[\mathbf{E}|\mathbf{c}]$. If a row of the form

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & | & \alpha \end{pmatrix}, \quad \alpha \neq 0$$

does not appear in $[\mathbf{E}|\mathbf{c}]$, is it possible that rows of this form could have appeared at earlier stages in the reduction process? Why?

2.4 HOMOGENEOUS SYSTEMS

A system of m linear equations in n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0, \end{aligned}$$

in which the right-hand side consists entirely of 0's is said to be a ***homogeneous system.*** If there is at least one nonzero number on the right-hand side, then the system is called ***nonhomogeneous.*** The purpose of this section is to examine some of the elementary aspects concerning homogeneous systems.

Consistency is never an issue when dealing with homogeneous systems because the zero solution $x_1 = x_2 = \cdots = x_n = 0$ is always one solution regardless of the values of the coefficients. Hereafter, the solution consisting of all zeros is referred to as the ***trivial solution.*** The only question is, "Are there solutions other than the trivial solution, and if so, how can we best describe them?" As before, Gaussian elimination provides the answer.

While reducing the augmented matrix $[\mathbf{A}|\mathbf{0}]$ of a homogeneous system to a row echelon form using Gaussian elimination, the zero column on the right-hand side can never be altered by any of the three elementary row operations. That is, any row echelon form derived from $[\mathbf{A}|\mathbf{0}]$ by means of row operations must also have the form $[\mathbf{E}|\mathbf{0}]$. This means that the last column of 0's is just excess baggage that is not necessary to carry along at each step. Just reduce the coefficient matrix $\mathbf{A}$ to a row echelon form $\mathbf{E}$, and remember that the right-hand side is entirely zero when you execute back substitution. The process is best understood by considering a typical example.

In order to examine the solutions of the homogeneous system

$$\begin{aligned} x_1 + 2x_2 + 2x_3 + 3x_4 &= 0, \\ 2x_1 + 4x_2 + x_3 + 3x_4 &= 0, \\ 3x_1 + 6x_2 + x_3 + 4x_4 &= 0, \end{aligned} \tag{2.4.1}$$

reduce the coefficient matrix to a row echelon form.

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

Therefore, the original homogeneous system is equivalent to the following reduced homogeneous system:

$$\begin{aligned} x_1 + 2x_2 + 2x_3 + 3x_4 &= 0, \\ -3x_3 - 3x_4 &= 0. \end{aligned} \tag{2.4.2}$$

Since there are four unknowns but only two equations in this reduced system, it is impossible to extract a unique solution for each unknown. The best we can do is to pick two "basic" unknowns—which will be called the ***basic variables*** and solve for these in terms of the other two unknowns—whose values must remain arbitrary or "free," and consequently they will be referred to as the ***free variables.*** Although there are several possibilities for selecting a set of basic variables, the convention is to *always solve for the unknowns corresponding to the pivotal positions*—or, equivalently, the unknowns corresponding to the basic columns. In this example, the pivots (as well as the basic columns) lie in the first and third positions, so the strategy is to apply back substitution to solve the reduced system (2.4.2) for the basic variables x_1 and x_3 in terms of the free variables x_2 and x_4. The second equation in (2.4.2) yields

$$x_3 = -x_4$$

and substitution back into the first equation produces

$$\begin{aligned} x_1 &= -2x_2 - 2x_3 - 3x_4, \\ &= -2x_2 - 2(-x_4) - 3x_4, \\ &= -2x_2 - x_4. \end{aligned}$$

Therefore, all solutions of the original homogeneous system can be described by saying

$$\begin{aligned} &x_1 = -2x_2 - x_4, \\ &x_2 \text{ is "free,"} \\ &x_3 = -x_4, \\ &x_4 \text{ is "free."} \end{aligned} \tag{2.4.3}$$

As the free variables x_2 and x_4 range over all possible values, the above expressions describe all possible solutions. For example, when x_2 and x_4 assume the values $x_2 = 1$ and $x_4 = -2$, then the particular solution

$$x_1 = 0, \quad x_2 = 1, \quad x_3 = 2, \quad x_4 = -2$$

is produced. When $x_2 = \pi$ and $x_4 = \sqrt{2}$, then another particular solution

$$x_1 = -2\pi - \sqrt{2}, \quad x_2 = \pi, \quad x_3 = -\sqrt{2}, \quad x_4 = \sqrt{2}$$

is generated.

Rather than describing the solution set as illustrated in (2.4.3), future developments will make it more convenient to express the solution set by writing

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -2x_2 - x_4 \\ x_2 \\ -x_4 \\ x_4 \end{pmatrix} = x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix} \tag{2.4.4}$$

with the understanding that x_2 and x_4 are free variables that can range over all possible numbers. This representation will be called the ***general solution*** of the homogeneous system. This expression for the general solution emphasizes that every solution is some combination of the two particular solutions

$$\mathbf{h}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{h}_2 = \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}.$$

The fact that $\mathbf{h}_1$ and $\mathbf{h}_2$ are each solutions is clear because $\mathbf{h}_1$ is produced when the free variables assume the values $x_2 = 1$ and $x_4 = 0$, whereas the solution $\mathbf{h}_2$ is generated when $x_2 = 0$ and $x_4 = 1$.

Now consider a general homogeneous system $[\mathbf{A}|\mathbf{0}]$ of m linear equations in n unknowns. If the coefficient matrix is such that $rank\,(\mathbf{A}) = r$, then it should be apparent from the preceding discussion that there will be exactly r basic variables—corresponding to the positions of the basic columns in $\mathbf{A}$—and exactly $n - r$ free variables—corresponding to the positions of the nonbasic columns in $\mathbf{A}$. Reducing $\mathbf{A}$ to a row echelon form using Gaussian elimination and then using back substitution to solve for the basic variables in terms of the free variables produces the ***general solution,*** which has the form

$$\mathbf{x} = x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}, \tag{2.4.5}$$

where $x_{f_1}, x_{f_2}, \ldots, x_{f_{n-r}}$ are the free variables and where $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}$ are $n \times 1$ columns that represent particular solutions of the system. As the free variables x_{f_i} range over all possible values, the general solution generates all possible solutions.

The general solution does not depend on which row echelon form is used in the sense that using back substitution to solve for the basic variables in terms of the nonbasic variables generates a unique set of particular solutions $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}\}$, regardless of which row echelon form is used. Without going into great detail, one can argue that this is true because using back substitution in any row echelon form to solve for the basic variables must produce exactly the same result as that obtained by completely reducing $\mathbf{A}$ to $\mathbf{E_A}$ and then solving the reduced homogeneous system for the basic variables. Uniqueness of $\mathbf{E_A}$ guarantees the uniqueness of the $\mathbf{h}_i$'s.

For example, if the coefficient matrix $\mathbf{A}$ associated with the system (2.4.1) is completely reduced by the Gauss–Jordan procedure to $\mathbf{E_A}$

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 & 3 \\ 2 & 4 & 1 & 3 \\ 3 & 6 & 1 & 4 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E_A},$$

then we obtain the following reduced system:

$$\begin{aligned} x_1 + 2x_2 + x_4 &= 0, \\ x_3 + x_4 &= 0. \end{aligned}$$

Solving for the basic variables x_1 and x_3 in terms of x_2 and x_4 produces exactly the same result as given in (2.4.3) and hence generates exactly the same general solution as shown in (2.4.4).

Because it avoids the back substitution process, you may find it more convenient to use the Gauss–Jordan procedure to reduce $\mathbf{A}$ completely to $\mathbf{E_A}$ and then construct the general solution directly from the entries in $\mathbf{E_A}$. This approach usually will be adopted in the examples and exercises.

As was previously observed, all homogeneous systems are consistent because the trivial solution consisting of all zeros is always one solution. The natural question is, "When is the trivial solution the *only* solution?" In other words, we wish to know when a homogeneous system possesses a unique solution. The form of the general solution (2.4.5) makes the answer transparent. As long as there is at least one free variable, then it is clear from (2.4.5) that there will be an infinite number of solutions. Consequently, the trivial solution is the only solution if and only if there are no free variables. Because the number of free variables is given by $n - r$, where $r = rank\,(\mathbf{A})$, the previous statement can be reformulated to say that *a homogeneous system possesses a unique solution—the trivial solution—if and only if* $rank\,(\mathbf{A}) = n$.

Example 2.4.1

The homogeneous system

$$\begin{aligned} x_1 + 2x_2 + 2x_3 &= 0, \\ 2x_1 + 5x_2 + 7x_3 &= 0, \\ 3x_1 + 6x_2 + 8x_3 &= 0, \end{aligned}$$

has only the trivial solution because

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 5 & 7 \\ 3 & 6 & 8 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{pmatrix} = \mathbf{E}$$

shows that $rank\,(\mathbf{A}) = n = 3$. Indeed, it is also obvious from $\mathbf{E}$ that applying back substitution in the system $[\mathbf{E}|\mathbf{0}]$ yields only the trivial solution.

Example 2.4.2

Problem: Explain why the following homogeneous system has infinitely many solutions, and exhibit the general solution:

$$\begin{aligned} x_1 + 2x_2 + 2x_3 &= 0, \\ 2x_1 + 5x_2 + 7x_3 &= 0, \\ 3x_1 + 6x_2 + 6x_3 &= 0. \end{aligned}$$

Solution:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 5 & 7 \\ 3 & 6 & 6 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{E}$$

shows that $rank\,(\mathbf{A}) = 2 < n = 3$. Since the basic columns lie in positions one and two, x_1 and x_2 are the basic variables while x_3 is free. Using back substitution on $[\mathbf{E}|\mathbf{0}]$ to solve for the basic variables in terms of the free variable produces $x_2 = -3x_3$ and $x_1 = -2x_2 - 2x_3 = 4x_3$, so the general solution is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_3 \begin{pmatrix} 4 \\ -3 \\ 1 \end{pmatrix}, \quad \text{where} \quad x_3 \text{ is free.}$$

That is, every solution is a multiple of the one particular solution $\mathbf{h}_1 = \begin{pmatrix} 4 \\ -3 \\ 1 \end{pmatrix}$.

Summary

Let $\mathbf{A}_{m\times n}$ be the coefficient matrix for a homogeneous system of m linear equations in n unknowns, and suppose $rank\,(\mathbf{A}) = r$.

- The unknowns that correspond to the positions of the basic columns (i.e., the pivotal positions) are called the ***basic variables***, and the unknowns corresponding to the positions of the nonbasic columns are called the ***free variables.***
- There are exactly r basic variables and $n - r$ free variables.
- To describe all solutions, reduce $\mathbf{A}$ to a row echelon form using Gaussian elimination, and then use back substitution to solve for the basic variables in terms of the free variables. This produces the ***general solution*** that has the form

  $$\mathbf{x} = x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r},$$

  where the terms $x_{f_1}, x_{f_2}, \ldots, x_{f_{n-r}}$ are the free variables and where $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}$ are $n \times 1$ columns that represent particular solutions of the homogeneous system. The $\mathbf{h}_i$'s are independent of which row echelon form is used in the back substitution process. As the free variables x_{f_i} range over all possible values, the general solution generates all possible solutions.
- A homogeneous system possesses a unique solution (the trivial solution) if and only if $rank\,(\mathbf{A}) = n$—i.e., if and only if there are no free variables.

Exercises for section 2.4

2.4.1. Determine the general solution for each of the following homogeneous systems.

(a) $$\begin{aligned} x_1 + 2x_2 + x_3 + 2x_4 &= 0, \\ 2x_1 + 4x_2 + x_3 + 3x_4 &= 0, \\ 3x_1 + 6x_2 + x_3 + 4x_4 &= 0. \end{aligned}$$

(b) $$\begin{aligned} 2x + y + z &= 0, \\ 4x + 2y + z &= 0, \\ 6x + 3y + z &= 0, \\ 8x + 4y + z &= 0. \end{aligned}$$

(c) $$\begin{aligned} x_1 + x_2 + 2x_3 \phantom{{}+ 3x_4} &= 0, \\ 3x_1 \phantom{{}+ x_2} + 3x_3 + 3x_4 &= 0, \\ 2x_1 + x_2 + 3x_3 + x_4 &= 0, \\ x_1 + 2x_2 + 3x_3 - x_4 &= 0. \end{aligned}$$

(d) $$\begin{aligned} 2x + y + z &= 0, \\ 4x + 2y + z &= 0, \\ 6x + 3y + z &= 0, \\ 8x + 5y + z &= 0. \end{aligned}$$

2.4.2. Among all solutions that satisfy the homogeneous system

$$\begin{aligned} x + 2y + z &= 0, \\ 2x + 4y + z &= 0, \\ x + 2y - z &= 0, \end{aligned}$$

determine those that also satisfy the nonlinear constraint $y - xy = 2z$.

2.4.3. Consider a homogeneous system whose coefficient matrix is

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

First transform $\mathbf{A}$ to an unreduced row echelon form to determine the general solution of the associated homogeneous system. Then reduce $\mathbf{A}$ to $\mathbf{E_A}$, and show that the same general solution is produced.

2.4.4. If $\mathbf{A}$ is the coefficient matrix for a homogeneous system consisting of four equations in eight unknowns and if there are five free variables, what is $rank\,(\mathbf{A})$?

2.4.5. Suppose that $\mathbf{A}$ is the coefficient matrix for a homogeneous system of four equations in six unknowns and suppose that $\mathbf{A}$ has at least one nonzero row.

(a) Determine the fewest number of free variables that are possible.

(b) Determine the maximum number of free variables that are possible.

2.4.6. Explain why a homogeneous system of m equations in n unknowns where $m < n$ must always possess an infinite number of solutions.

2.4.7. Construct a homogeneous system of three equations in four unknowns that has

$$x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -3 \\ 0 \\ 2 \\ 1 \end{pmatrix}$$

as its general solution.

2.4.8. If $\mathbf{c}_1$ and $\mathbf{c}_2$ are columns that represent two particular solutions of the same homogeneous system, explain why the sum $\mathbf{c}_1 + \mathbf{c}_2$ must also represent a solution of this system.

2.5 NONHOMOGENEOUS SYSTEMS

Recall that a system of m linear equations in n unknowns

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m, \end{array}$$

is said to be ***nonhomogeneous*** whenever $b_i \neq 0$ for at least one i. Unlike homogeneous systems, a nonhomogeneous system may be inconsistent and the techniques of §2.3 must be applied in order to determine if solutions do indeed exist. Unless otherwise stated, it is assumed that all systems in this section are consistent.

To describe the set of all possible solutions of a consistent nonhomogeneous system, construct a general solution by exactly the same method used for homogeneous systems as follows.

- Use Gaussian elimination to reduce the associated augmented matrix $[\mathbf{A}|\mathbf{b}]$ to a row echelon form $[\mathbf{E}|\mathbf{c}]$.

- Identify the basic variables and the free variables in the same manner described in §2.4.

- Apply back substitution to $[\mathbf{E}|\mathbf{c}]$ and solve for the basic variables in terms of the free variables.

- Write the result in the form

$$\mathbf{x} = \mathbf{p} + x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}, \tag{2.5.1}$$

  where $x_{f_1}, x_{f_2}, \ldots, x_{f_{n-r}}$ are the free variables and $\mathbf{p}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}$ are $n \times 1$ columns. This is the ***general solution*** of the nonhomogeneous system.

As the free variables x_{f_i} range over all possible values, the general solution (2.5.1) generates all possible solutions of the system $[\mathbf{A}|\mathbf{b}]$. Just as in the homogeneous case, the columns $\mathbf{h_i}$ and $\mathbf{p}$ are independent of which row echelon form $[\mathbf{E}|\mathbf{c}]$ is used. Therefore, $[\mathbf{A}|\mathbf{b}]$ may be completely reduced to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$ by using the Gauss–Jordan method thereby avoiding the need to perform back substitution. We will use this approach whenever it is convenient.

The difference between the general solution of a nonhomogeneous system and the general solution of a homogeneous system is the column $\mathbf{p}$ that appears

in (2.5.1). To understand why $\mathbf{p}$ appears and where it comes from, consider the nonhomogeneous system

$$\begin{aligned} x_1 + 2x_2 + 2x_3 + 3x_4 &= 4, \\ 2x_1 + 4x_2 + x_3 + 3x_4 &= 5, \\ 3x_1 + 6x_2 + x_3 + 4x_4 &= 7, \end{aligned} \tag{2.5.2}$$

in which the coefficient matrix is the same as the coefficient matrix for the homogeneous system (2.4.1) used in the previous section. If $[\mathbf{A}|\mathbf{b}]$ is completely reduced by the Gauss–Jordan procedure to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$

$$[\mathbf{A}|\mathbf{b}] = \left(\begin{array}{cccc|c} 1 & 2 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 & 5 \\ 3 & 6 & 1 & 4 & 7 \end{array}\right) \longrightarrow \left(\begin{array}{cccc|c} 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right) = \mathbf{E}_{[\mathbf{A}|\mathbf{b}]},$$

then the following reduced system is obtained:

$$\begin{aligned} x_1 + 2x_2 + x_4 &= 2, \\ x_3 + x_4 &= 1. \end{aligned}$$

Solving for the basic variables, x_1 and x_3, in terms of the free variables, x_2 and x_4, produces

$$\begin{aligned} &x_1 = 2 - 2x_2 - x_4, \\ &x_2 \text{ is "free,"} \\ &x_3 = 1 - x_4, \\ &x_4 \text{ is "free."} \end{aligned}$$

The general solution is obtained by writing these statements in the form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 - 2x_2 - x_4 \\ x_2 \\ 1 - x_4 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}. \tag{2.5.3}$$

As the free variables x_2 and x_4 range over all possible numbers, this generates all possible solutions of the nonhomogeneous system (2.5.2). Notice that the column $\begin{pmatrix} 2 \\ 0 \\ 1 \\ 0 \end{pmatrix}$ in (2.5.3) is a *particular solution* of the nonhomogeneous system (2.5.2)—it is the solution produced when the free variables assume the values $x_2 = 0$ and $x_4 = 0$.

Furthermore, recall from (2.4.4) that the general solution of the associated homogeneous system

$$\begin{aligned} x_1 + 2x_2 + 2x_3 + 3x_4 &= 0, \\ 2x_1 + 4x_2 + x_3 + 3x_4 &= 0, \\ 3x_1 + 6x_2 + x_3 + 4x_4 &= 0, \end{aligned} \tag{2.5.4}$$

is given by

$$\begin{pmatrix} -2x_2 - x_4 \\ x_2 \\ -x_4 \\ x_4 \end{pmatrix} = x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}.$$

That is, the general solution of the associated homogeneous system (2.5.4) is a part of the general solution of the original nonhomogeneous system (2.5.2).

These two observations can be combined by saying that *the general solution of the nonhomogeneous system is given by a particular solution plus the general solution of the associated homogeneous system.*[14]

To see that the previous statement is always true, suppose $[\mathbf{A}|\mathbf{b}]$ represents a general $m \times n$ consistent system where $rank\,(\mathbf{A}) = r$. Consistency guarantees that $\mathbf{b}$ is a nonbasic column in $[\mathbf{A}|\mathbf{b}]$, and hence the basic columns in $[\mathbf{A}|\mathbf{b}]$ are in the same positions as the basic columns in $[\mathbf{A}|\mathbf{0}]$ so that the nonhomogeneous system and the associated homogeneous system have exactly the same set of basic variables as well as free variables. Furthermore, it is not difficult to see that

$$\mathbf{E}_{[\mathbf{A}|\mathbf{0}]} = [\mathbf{E}_{\mathbf{A}}|\mathbf{0}] \quad \text{and} \quad \mathbf{E}_{[\mathbf{A}|\mathbf{b}]} = [\mathbf{E}_{\mathbf{A}}|\mathbf{c}],$$

where $\mathbf{c}$ is some column of the form $\mathbf{c} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_r \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. This means that if you solve the ith equation in the reduced homogeneous system for the ith basic variable x_{b_i} in terms of the free variables $x_{f_i}, x_{f_{i+1}}, \ldots, x_{f_{n-r}}$ to produce

$$x_{b_i} = \alpha_i x_{f_i} + \alpha_{i+1} x_{f_{i+1}} + \cdots + \alpha_{n-r} x_{f_{n-r}}, \tag{2.5.5}$$

then the solution for the ith basic variable in the reduced nonhomogeneous system must have the form

$$x_{b_i} = \xi_i + \alpha_i x_{f_i} + \alpha_{i+1} x_{f_{i+1}} + \cdots + \alpha_{n-r} x_{f_{n-r}}. \tag{2.5.6}$$

[14] For those students who have studied differential equations, this statement should have a familiar ring. Exactly the same situation holds for the general solution to a linear differential equation. This is no accident—it is due to the inherent linearity in both problems. More will be said about this issue later in the text.

That is, the two solutions differ only in the fact that the latter contains the constant ξ_i. Consider organizing the expressions (2.5.5) and (2.5.6) so as to construct the respective general solutions. If the general solution of the homogeneous system has the form

$$\mathbf{x} = x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r},$$

then it is apparent that the general solution of the nonhomogeneous system must have a similar form

$$\mathbf{x} = \mathbf{p} + x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r} \tag{2.5.7}$$

in which the column $\mathbf{p}$ contains the constants ξ_i along with some 0's—the ξ_i's occupy positions in $\mathbf{p}$ that correspond to the positions of the basic columns, and 0's occupy all other positions. The column $\mathbf{p}$ represents one particular solution to the nonhomogeneous system because it is the solution produced when the free variables assume the values $x_{f_1} = x_{f_2} = \cdots = x_{f_{n-r}} = 0$.

Example 2.5.1

Problem: Determine the general solution of the following nonhomogeneous system and compare it with the general solution of the associated homogeneous system:

$$\begin{aligned} x_1 + x_2 + 2x_3 + 2x_4 + x_5 &= 1,\\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 3x_5 &= 1,\\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 2x_5 &= 2,\\ 3x_1 + 5x_2 + 8x_3 + 6x_4 + 5x_5 &= 3. \end{aligned}$$

Solution: Reducing the augmented matrix $[\mathbf{A}|\mathbf{b}]$ to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$ yields

$$\mathbf{A} = \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1\\ 2 & 2 & 4 & 4 & 3 & 1\\ 2 & 2 & 4 & 4 & 2 & 2\\ 3 & 5 & 8 & 6 & 5 & 3 \end{array}\right) \longrightarrow \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1\\ 0 & 0 & 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 2 & 2 & 0 & 2 & 0 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1\\ 0 & 2 & 2 & 0 & 2 & 0\\ 0 & 0 & 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \longrightarrow \left(\begin{array}{ccccc|c} 1 & 1 & 2 & 2 & 1 & 1\\ 0 & 1 & 1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 2 & 0 & 1\\ 0 & 1 & 1 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \longrightarrow \left(\begin{array}{ccccc|c} 1 & 0 & 1 & 2 & 0 & 1\\ 0 & 1 & 1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 1 & -1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) = \mathbf{E}_{[\mathbf{A}|\mathbf{b}]}.$$

Observe that the system is indeed consistent because the last column is nonbasic. Solve the reduced system for the basic variables x_1, x_2, and x_5 in terms of the free variables x_3 and x_4 to obtain

$$\begin{aligned} x_1 &= 1 - x_3 - 2x_4, \\ x_2 &= 1 - x_3, \\ x_3 &\text{ is "free,"} \\ x_4 &\text{ is "free,"} \\ x_5 &= -1. \end{aligned}$$

The general solution to the nonhomogeneous system is

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 - x_3 - 2x_4 \\ 1 - x_3 \\ x_3 \\ x_4 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ -1 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -2 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

The general solution of the associated homogeneous system is

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -x_3 - 2x_4 \\ -x_3 \\ x_3 \\ x_4 \\ 0 \end{pmatrix} = x_3 \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -2 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

You should verify for yourself that

$$\mathbf{p} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

is indeed a particular solution to the nonhomogeneous system and that

$$\mathbf{h}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{h}_4 = \begin{pmatrix} -2 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

are particular solutions to the associated homogeneous system.

Now turn to the question, "When does a consistent system have a unique solution?" It is known from (2.5.7) that the general solution of a consistent $m \times n$ nonhomogeneous system $[\mathbf{A}|\mathbf{b}]$ with $rank\,(\mathbf{A}) = r$ is given by

$$\mathbf{x} = \mathbf{p} + x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r},$$

where

$$x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}$$

is the general solution of the associated homogeneous system. Consequently, it is evident that the nonhomogeneous system $[\mathbf{A}|\mathbf{b}]$ will have a unique solution (namely, $\mathbf{p}$) if and only if there are no free variables—i.e., if and only if $r = n$ (= number of unknowns)—this is equivalent to saying that the associated homogeneous system $[\mathbf{A}|\mathbf{0}]$ has only the trivial solution.

Example 2.5.2

Consider the following nonhomogeneous system:

$$\begin{aligned} 2x_1 + 4x_2 + 6x_3 &= 2, \\ x_1 + 2x_2 + 3x_3 &= 1, \\ x_1 \qquad\quad + \ x_3 &= -3, \\ 2x_1 + 4x_2 \qquad\quad &= 8. \end{aligned}$$

Reducing $[\mathbf{A}|\mathbf{b}]$ to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$ yields

$$[\mathbf{A}|\mathbf{b}] = \left(\begin{array}{ccc|c} 2 & 4 & 6 & 2 \\ 1 & 2 & 3 & 1 \\ 1 & 0 & 1 & -3 \\ 2 & 4 & 0 & 8 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & -2 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right) = \mathbf{E}_{[\mathbf{A}|\mathbf{b}]}.$$

The system is consistent because the last column is nonbasic. There are several ways to see that the system has a unique solution. Notice that

$$rank\,(\mathbf{A}) = 3 = \text{ number of unknowns},$$

which is the same as observing that there are no free variables. Furthermore, the associated homogeneous system clearly has only the trivial solution. Finally, because we completely reduced $[\mathbf{A}|\mathbf{b}]$ to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$, it is obvious that there is only one solution possible and that it is given by $\mathbf{p} = \begin{pmatrix} -2 \\ 3 \\ -1 \end{pmatrix}$.

Summary

Let $[\mathbf{A}|\mathbf{b}]$ be the augmented matrix for a consistent $m \times n$ nonhomogeneous system in which $rank\,(\mathbf{A}) = r$.

- Reducing $[\mathbf{A}|\mathbf{b}]$ to a row echelon form using Gaussian elimination and then solving for the basic variables in terms of the free variables leads to the ***general solution***

$$\mathbf{x} = \mathbf{p} + x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}.$$

  As the free variables x_{f_i} range over all possible values, this general solution generates all possible solutions of the system.
- Column $\mathbf{p}$ is a particular solution of the nonhomogeneous system.
- The expression $x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}$ is the general solution of the associated homogeneous system.
- Column $\mathbf{p}$ as well as the columns $\mathbf{h}_i$ are independent of the row echelon form to which $[\mathbf{A}|\mathbf{b}]$ is reduced.
- The system possesses a unique solution if and only if any of the following is true.
  - ▷ $rank\,(\mathbf{A}) = n =$ number of unknowns.
  - ▷ There are no free variables.
  - ▷ The associated homogeneous system possesses only the trivial solution.

Exercises for section 2.5

2.5.1. Determine the general solution for each of the following nonhomogeneous systems.

(a) $$\begin{aligned} x_1 + 2x_2 + x_3 + 2x_4 &= 3, \\ 2x_1 + 4x_2 + x_3 + 3x_4 &= 4, \\ 3x_1 + 6x_2 + x_3 + 4x_4 &= 5. \end{aligned}$$

(b) $$\begin{aligned} 2x + y + z &= 4, \\ 4x + 2y + z &= 6, \\ 6x + 3y + z &= 8, \\ 8x + 4y + z &= 10. \end{aligned}$$

(c) $$\begin{aligned} x_1 + x_2 + 2x_3 &= 1, \\ 3x_1 + 3x_3 + 3x_4 &= 6, \\ 2x_1 + x_2 + 3x_3 + x_4 &= 3, \\ x_1 + 2x_2 + 3x_3 - x_4 &= 0. \end{aligned}$$

(d) $$\begin{aligned} 2x + y + z &= 2, \\ 4x + 2y + z &= 5, \\ 6x + 3y + z &= 8, \\ 8x + 5y + z &= 8. \end{aligned}$$

2.5.2. Among the solutions that satisfy the set of linear equations

$$\begin{aligned} x_1 + x_2 + 2x_3 + 2x_4 + x_5 &= 1, \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 3x_5 &= 1, \\ 2x_1 + 2x_2 + 4x_3 + 4x_4 + 2x_5 &= 2, \\ 3x_1 + 5x_2 + 8x_3 + 6x_4 + 5x_5 &= 3, \end{aligned}$$

find all those that also satisfy the following two constraints:

$$\begin{aligned} (x_1 - x_2)^2 - 4x_5^2 &= 0, \\ x_3^2 - x_5^2 &= 0. \end{aligned}$$

2.5.3. In order to grow a certain crop, it is recommended that each square foot of ground be treated with 10 units of phosphorous, 9 units of potassium, and 19 units of nitrogen. Suppose that there are three brands of fertilizer on the market—say brand $\mathcal{X}$, brand $\mathcal{Y}$, and brand $\mathcal{Z}$. One pound of brand $\mathcal{X}$ contains 2 units of phosphorous, 3 units of potassium, and 5 units of nitrogen. One pound of brand $\mathcal{Y}$ contains 1 unit of phosphorous, 3 units of potassium, and 4 units of nitrogen. One pound of brand $\mathcal{Z}$ contains only 1 unit of phosphorous and 1 unit of nitrogen.

(a) Take into account the obvious fact that a negative number of pounds of any brand can never be applied, and suppose that because of the way fertilizer is sold only an integral number of pounds of each brand will be applied. Under these constraints, determine all possible combinations of the three brands that can be applied to satisfy the recommendations exactly.

(b) Suppose that brand $\mathcal{X}$ costs \$1 per pound, brand $\mathcal{Y}$ costs \$6 per pound, and brand $\mathcal{Z}$ costs \$3 per pound. Determine the least expensive solution that will satisfy the recommendations exactly as well as the constraints of part (a).

2.5.4. Consider the following system:

$$\begin{aligned} 2x + 2y + 3z &= 0, \\ 4x + 8y + 12z &= -4, \\ 6x + 2y + \alpha z &= 4. \end{aligned}$$

(a) Determine all values of α for which the system is consistent.

(b) Determine all values of α for which there is a unique solution, and compute the solution for these cases.

(c) Determine all values of α for which there are infinitely many different solutions, and give the general solution for these cases.

2.5.5. If columns $\mathbf{s}_1$ and $\mathbf{s}_2$ are particular solutions of the same nonhomogeneous system, must it be the case that the sum $\mathbf{s}_1 + \mathbf{s}_2$ is also a solution?

2.5.6. Suppose that $[\mathbf{A}|\mathbf{b}]$ is the augmented matrix for a consistent system of m equations in n unknowns where $m \geq n$. What must $\mathbf{E_A}$ look like when the system possesses a unique solution?

2.5.7. Construct a nonhomogeneous system of three equations in four unknowns that has

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -3 \\ 0 \\ 2 \\ 1 \end{pmatrix}$$

as its general solution.

2.5.8. Consider using floating-point arithmetic (without partial pivoting or scaling) to solve the system represented by the following augmented matrix:

$$\left(\begin{array}{ccc|c} .835 & .667 & .5 & .168 \\ .333 & .266 & .1994 & .067 \\ 1.67 & 1.334 & 1.1 & .436 \end{array}\right).$$

(a) Determine the 4-digit general solution.
(b) Determine the 5-digit general solution.
(c) Determine the 6-digit general solution.

2.6 ELECTRICAL CIRCUITS

The theory of electrical circuits is an important application that naturally gives rise to rectangular systems of linear equations. Because the underlying mathematics depends on several of the concepts discussed in the preceding sections, you may find it interesting and worthwhile to make a small excursion into the elementary mathematical analysis of electrical circuits. However, the continuity of the text is not compromised by omitting this section.

In a direct current circuit containing resistances and sources of electromotive force (abbreviated EMF) such as batteries, a point at which three or more conductors are joined is called a ***node*** or ***branch point*** of the circuit, and a closed conduction path is called a ***loop.*** Any part of a circuit between two adjoining nodes is called a ***branch*** of the circuit. The circuit shown in Figure 2.6.1 is a typical example that contains four nodes, seven loops, and six branches.

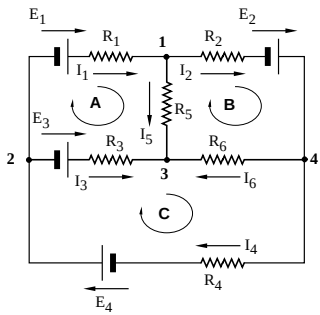


FIGURE 2.6.1

The problem is to relate the currents I_k in each branch to the resistances R_k and the EMFs E_k.[15] This is accomplished by using ***Ohm's law*** in conjunction with ***Kirchhoff's rules*** to produce a system of linear equations.

Ohm's Law

Ohm's law states that for a current of I amps, the voltage drop (in volts) across a resistance of R ohms is given by $V = IR$.

Kirchhoff's rules—formally stated below—are the two fundamental laws that govern the study of electrical circuits.

[15] For an EMF source of magnitude E and a current I, there is always a small internal resistance in the source, and the voltage drop across it is $V = E - I \times (\textit{internal resistance})$. But internal source resistance is usually negligible, so the voltage drop across the source can be taken as $V = E$. When internal resistance cannot be ignored, its effects may be incorporated into existing external resistances, or it can be treated as a separate external resistance.

Kirchhoff's Rules

NODE RULE: *The algebraic sum of currents toward each node is zero.* That is, the total incoming current must equal the total outgoing current. This is simply a statement of conservation of charge.

LOOP RULE: *The algebraic sum of the EMFs around each loop must equal the algebraic sum of the IR products in the same loop.* That is, assuming internal source resistances have been accounted for, the algebraic sum of the voltage drops over the sources equals the algebraic sum of the voltage drops over the resistances in each loop. This is a statement of conservation of energy.

Kirchhoff's rules may be used without knowing the directions of the currents and EMFs in advance. You may arbitrarily assign directions. If negative values emerge in the final solution, then the actual direction is opposite to that assumed. To apply the node rule, consider a current to be *positive* if its direction is toward the node—otherwise, consider the current to be *negative.* It should be clear that the node rule will always generate a homogeneous system. For example, applying the node rule to the circuit in Figure 2.6.1 yields four homogeneous equations in six unknowns—the unknowns are the I_k's:

$$\begin{aligned}
\text{Node 1:} \quad & I_1 - I_2 - I_5 = 0, \\
\text{Node 2:} \quad & -I_1 - I_3 + I_4 = 0, \\
\text{Node 3:} \quad & I_3 + I_5 + I_6 = 0, \\
\text{Node 4:} \quad & I_2 - I_4 - I_6 = 0.
\end{aligned}$$

To apply the loop rule, some direction (clockwise or counterclockwise) must be chosen as the positive direction, and all EMFs and currents in that direction are considered *positive* and those in the opposite direction are *negative.* It is possible for a current to be considered positive for the node rule but considered negative when it is used in the loop rule. If the positive direction is considered to be clockwise in each case, then applying the loop rule to the three indicated loops A, B, and C in the circuit shown in Figure 2.6.1 produces the three nonhomogeneous equations in six unknowns—the I_k's are treated as the unknowns, while the R_k's and E_k's are assumed to be known.

$$\begin{aligned}
\text{Loop A:} \quad & I_1R_1 - I_3R_3 + I_5R_5 = E_1 - E_3, \\
\text{Loop B:} \quad & I_2R_2 - I_5R_5 + I_6R_6 = E_2, \\
\text{Loop C:} \quad & I_3R_3 + I_4R_4 - I_6R_6 = E_3 + E_4.
\end{aligned}$$

There are 4 additional loops that also produce loop equations thereby making a total of 11 equations (4 nodal equations and 7 loop equations) in 6 unknowns. Although this appears to be a rather general 11×6 system of equations, it really is not. If the circuit is in a state of equilibrium, then the physics of the situation dictates that for each set of EMFs E_k, the corresponding currents I_k must be uniquely determined. In other words, physics guarantees that the 11×6 system produced by applying the two Kirchhoff rules must be *consistent* and possess a *unique* solution.

Suppose that $[\mathbf{A}|\mathbf{b}]$ represents the augmented matrix for the 11×6 system generated by Kirchhoff's rules. From the results in §2.5, we know that the system has a unique solution if and only if

$$rank\,(\mathbf{A}) = \text{ number of unknowns } = 6.$$

Furthermore, it was demonstrated in §2.3 that the system is consistent if and only if

$$rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A}).$$

Combining these two facts allows us to conclude that

$$rank[\mathbf{A}|\mathbf{b}] = 6$$

so that when $[\mathbf{A}|\mathbf{b}]$ is reduced to $\mathbf{E}_{[\mathbf{A}|\mathbf{b}]}$, there will be exactly 6 nonzero rows and 5 zero rows. Therefore, 5 of the original 11 equations are redundant in the sense that they can be "zeroed out" by forming combinations of some particular set of 6 "independent" equations. It is desirable to know beforehand which of the 11 equations will be redundant and which can act as the "independent" set.

Notice that in using the node rule, the equation corresponding to node 4 is simply the negative sum of the equations for nodes 1, 2, and 3, and that the first three equations are independent in the sense that no one of the three can be written as a combination of any other two. This situation is typical. For a general circuit with n nodes, it can be demonstrated that the equations for the first $n-1$ nodes are independent, and the equation for the last node is redundant.

The loop rule also can generate redundant equations. Only ***simple loops***—loops not containing smaller loops—give rise to independent equations. For example, consider the loop consisting of the three exterior branches in the circuit shown in Figure 2.6.1. Applying the loop rule to this large loop will produce no new information because the large loop can be constructed by "adding" the three simple loops A, B, and C contained within. The equation associated with the large outside loop is

$$I_1R_1 + I_2R_2 + I_4R_4 = E_1 + E_2 + E_4,$$

which is precisely the sum of the equations that correspond to the three component loops A, B, and C. This phenomenon will hold in general so that only the simple loops need to be considered when using the loop rule.

The point of this discussion is to conclude that the more general 11×6 rectangular system can be replaced by an equivalent 6×6 square system that has a unique solution by dropping the last nodal equation and using only the simple loop equations. This is characteristic of practical work in general. The physics of a problem together with natural constraints can usually be employed to replace a general rectangular system with one that is square and possesses a unique solution.

One of the goals in our study is to understand more clearly the notion of "independence" that emerged in this application. So far, independence has been an intuitive idea, but this example helps make it clear that independence is a fundamentally important concept that deserves to be nailed down more firmly. This is done in §4.3, and the general theory for obtaining independent equations from electrical circuits is developed in Examples 4.4.6 and 4.4.7.

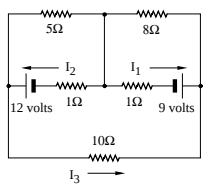
Exercises for section 2.6

2.6.1. Suppose that $R_i = i$ ohms and $E_i = i$ volts in the circuit shown in Figure 2.6.1.

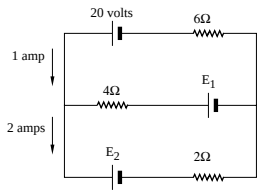
(a) Determine the six indicated currents.

(b) Select node number 1 to use as a reference point and fix its potential to be 0 volts. With respect to this reference, calculate the potentials at the other three nodes. Check your answer by verifying the loop rule for each loop in the circuit.

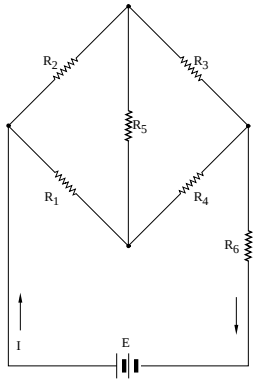
2.6.2. Determine the three currents indicated in the following circuit.



2.6.3. Determine the two unknown EMFs in the following circuit.



2.6.4. Consider the circuit shown below and answer the following questions.



(a) How many nodes does the circuit contain?
(b) How many branches does the circuit contain?
(c) Determine the total number of loops and then determine the number of simple loops.
(d) Demonstrate that the simple loop equations form an "independent" system of equations in the sense that there are no redundant equations.
(e) Verify that any three of the nodal equations constitute an "independent" system of equations.
(f) Verify that the loop equation associated with the loop containing R_1, R_2, R_3, and R_4 can be expressed as the sum of the two equations associated with the two simple loops contained in the larger loop.
(g) Determine the indicated current I if $R_1 = R_2 = R_3 = R_4 = 1$ ohm, $R_5 = R_6 = 5$ ohms, and $E = 5$ volts.

CHAPTER 3

Matrix Algebra

3.1 FROM ANCIENT CHINA TO ARTHUR CAYLEY

The ancient Chinese appreciated the advantages of array manipulation in dealing with systems of linear equations, and they possessed the seed that might have germinated into a genuine theory of matrices. Unfortunately, in the year 213 B.C., emperor Shih Hoang-ti ordered that "all books be burned and all scholars be buried." It is presumed that the emperor wanted all knowledge and written records to begin with him and his regime. The edict was carried out, and it will never be known how much knowledge was lost. The book *Chiu-chang Suan-shu* (*Nine Chapters on Arithmetic*), mentioned in the introduction to Chapter 1, was compiled on the basis of remnants that survived.

More than a millennium passed before further progress was documented. The Chinese counting board with its colored rods and its applications involving array manipulation to solve linear systems eventually found its way to Japan. Seki Kowa (1642–1708), whom many Japanese consider to be one of the greatest mathematicians that their country has produced, carried forward the Chinese principles involving "rule of thumb" elimination methods on arrays of numbers. His understanding of the elementary operations used in the Chinese elimination process led him to formulate the concept of what we now call the determinant. While formulating his ideas concerning the solution of linear systems, Seki Kowa anticipated the fundamental concepts of array operations that today form the basis for matrix algebra. However, there is no evidence that he developed his array operations to actually construct an algebra for matrices.

From the middle 1600s to the middle 1800s, while Europe was flowering in mathematical development, the study of array manipulation was exclusively

dedicated to the theory of determinants. Curiously, matrix algebra did not evolve along with the study of determinants.

It was not until the work of the British mathematician Arthur Cayley (1821–1895) that the matrix was singled out as a separate entity, distinct from the notion of a determinant, and algebraic operations between matrices were defined. In an 1855 paper, Cayley first introduced his basic ideas that were presented mainly to simplify notation. Finally, in 1857, Cayley expanded on his original ideas and wrote *A Memoir on the Theory of Matrices.* This laid the foundations for the modern theory and is generally credited for being the birth of the subjects of matrix analysis and linear algebra.

Arthur Cayley began his career by studying literature at Trinity College, Cambridge (1838–1842), but developed a side interest in mathematics, which he studied in his spare time. This "hobby" resulted in his first mathematical paper in 1841 when he was only 20 years old. To make a living, he entered the legal profession and practiced law for 14 years. However, his main interest was still mathematics. During the legal years alone, Cayley published almost 300 papers in mathematics.

In 1850 Cayley crossed paths with James J. Sylvester, and between the two of them matrix theory was born and nurtured. The two have been referred to as the "invariant twins." Although Cayley and Sylvester shared many mathematical interests, they were quite different people, especially in their approach to mathematics. Cayley had an insatiable hunger for the subject, and he read everything that he could lay his hands on. Sylvester, on the other hand, could not stand the sight of papers written by others. Cayley never forgot anything he had read or seen—he became a living encyclopedia. Sylvester, so it is said, would frequently fail to remember even his own theorems.

In 1863, Cayley was given a chair in mathematics at Cambridge University, and thereafter his mathematical output was enormous. Only Cauchy and Euler were as prolific. Cayley often said, "I really love my subject," and all indications substantiate that this was indeed the way he felt. He remained a working mathematician until his death at age 74.

Because the idea of the determinant preceded concepts of matrix algebra by at least two centuries, Morris Kline says in his book *Mathematical Thought from Ancient to Modern Times* that "the subject of matrix theory was well developed before it was created." This must have indeed been the case because immediately after the publication of Cayley's memoir, the subjects of matrix theory and linear algebra virtually exploded and quickly evolved into a discipline that now occupies a central position in applied mathematics.

3.2 ADDITION AND TRANSPOSITION

In the previous chapters, matrix language and notation were used simply to formulate some of the elementary concepts surrounding linear systems. The purpose now is to turn this language into a mathematical theory.[16]

Unless otherwise stated, a ***scalar*** is a complex number. Real numbers are a subset of the complex numbers, and hence real numbers are also scalar quantities. In the early stages, there is little harm in thinking only in terms of real scalars. Later on, however, the necessity for dealing with complex numbers will be unavoidable. Throughout the text, $\Re$ will denote the set of real numbers, and $\mathcal{C}$ will denote the complex numbers. The set of all n-tuples of real numbers will be denoted by $\Re^n$, and the set of all complex n-tuples will be denoted by $\mathcal{C}^n$. For example, $\Re^2$ is the set of all ordered pairs of real numbers (i.e., the standard cartesian plane), and $\Re^3$ is ordinary 3-space. Analogously, $\Re^{m\times n}$ and $\mathcal{C}^{m\times n}$ denote the $m \times n$ matrices containing real numbers and complex numbers, respectively.

Matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$ are defined to be ***equal matrices*** when $\mathbf{A}$ and $\mathbf{B}$ have the same shape and corresponding entries are equal. That is, $a_{ij} = b_{ij}$ for each $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. In particular, this definition applies to arrays such as $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$. Even though $\mathbf{u}$ and $\mathbf{v}$ describe exactly the same point in 3-space, we cannot consider them to be equal matrices because they have different shapes. An array (or matrix) consisting of a single column, such as $\mathbf{u}$, is called a ***column vector***, while an array consisting of a single row, such as $\mathbf{v}$, is called a ***row vector***.

Addition of Matrices

If $\mathbf{A}$ and $\mathbf{B}$ are $m \times n$ matrices, the ***sum*** of $\mathbf{A}$ and $\mathbf{B}$ is defined to be the $m \times n$ matrix $\mathbf{A}+\mathbf{B}$ obtained by adding corresponding entries. That is,

$$[\mathbf{A}+\mathbf{B}]_{ij} = [\mathbf{A}]_{ij} + [\mathbf{B}]_{ij} \quad \text{for each } i \text{ and } j.$$

For example,

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

[16] The great French mathematician Pierre-Simon Laplace (1749–1827) said that, "Such is the advantage of a well-constructed language that its simplified notation often becomes the source of profound theories." The theory of matrices is a testament to the validity of Laplace's statement.

The symbol "+" is used two different ways—it denotes addition between scalars in some places and addition between matrices at other places. Although these are two distinct algebraic operations, no ambiguities will arise if the context in which "+" appears is observed. Also note that the requirement that $\mathbf{A}$ and $\mathbf{B}$ have the same shape prevents adding a row to a column, even though the two may contain the same number of entries.

The matrix $(-\mathbf{A})$, called the ***additive inverse*** of $\mathbf{A}$, is defined to be the matrix obtained by negating each entry of $\mathbf{A}$. That is, if $\mathbf{A} = [a_{ij}]$, then $-\mathbf{A} = [-a_{ij}]$. This allows matrix subtraction to be defined in the natural way. For two matrices of the same shape, the ***difference*** $\mathbf{A} - \mathbf{B}$ is defined to be the matrix $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$ so that

$$[\mathbf{A} - \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} - [\mathbf{B}]_{ij} \quad \text{for each } i \text{ and } j.$$

Since matrix addition is defined in terms of scalar addition, the familiar algebraic properties of scalar addition are inherited by matrix addition as detailed below.

Properties of Matrix Addition

For $m \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, the following properties hold.

Closure property:	$\mathbf{A} + \mathbf{B}$ is again an $m \times n$ matrix.
Associative property:	$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.
Commutative property:	$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
Additive identity:	The $m \times n$ matrix $\mathbf{0}$ consisting of all zeros has the property that $\mathbf{A} + \mathbf{0} = \mathbf{A}$.
Additive inverse:	The $m \times n$ matrix $(-\mathbf{A})$ has the property that $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$.

Another simple operation that is derived from scalar arithmetic is as follows.

Scalar Multiplication

The product of a scalar α times a matrix $\mathbf{A}$, denoted by $\alpha\mathbf{A}$, is defined to be the matrix obtained by multiplying each entry of $\mathbf{A}$ by α. That is, $[\alpha\mathbf{A}]_{ij} = \alpha[\mathbf{A}]_{ij}$ for each i and j.

For example,

$$2\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 4 & 6 \\ 0 & 2 & 4 \\ 2 & 8 & 4 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 2 & 4 \\ 6 & 8 \\ 0 & 2 \end{pmatrix}.$$

The rules for combining addition and scalar multiplication are what you might suspect they should be. Some of the important ones are listed below.

Properties of Scalar Multiplication

For $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ and for scalars α and β, the following properties hold.

Closure property: $\alpha\mathbf{A}$ is again an $m \times n$ matrix.

Associative property: $(\alpha\beta)\mathbf{A} = \alpha(\beta\mathbf{A})$.

Distributive property: $\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}$. Scalar multiplication is distributed over matrix addition.

Distributive property: $(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}$. Scalar multiplication is distributed over scalar addition.

Identity property: $1\mathbf{A} = \mathbf{A}$. The number 1 is an identity element under scalar multiplication.

Other properties such as $\alpha\mathbf{A} = \mathbf{A}\alpha$ could have been listed, but the properties singled out pave the way for the definition of a vector space on p. 160.

A matrix operation that's not derived from scalar arithmetic is *transposition* as defined below.

Transpose

The ***transpose*** of $\mathbf{A}_{m\times n}$ is defined to be the $n \times m$ matrix $\mathbf{A}^T$ obtained by interchanging rows and columns in $\mathbf{A}$. More precisely, if $\mathbf{A} = [a_{ij}]$, then $[\mathbf{A}^T]_{ij} = a_{ji}$. For example,

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

It should be evident that for all matrices, $\left(\mathbf{A}^T\right)^T = \mathbf{A}$.

Whenever a matrix contains complex entries, the operation of complex conjugation almost always accompanies the transpose operation. (Recall that the complex conjugate of $z = a + ib$ is defined to be $\overline{z} = a - ib$.)

Conjugate Transpose

For $\mathbf{A} = [a_{ij}]$, the ***conjugate matrix*** is defined to be $\overline{\mathbf{A}} = [\overline{a}_{ij}]$, and the ***conjugate transpose*** of $\mathbf{A}$ is defined to be $\bar{\mathbf{A}}^T = \overline{\mathbf{A}^T}$. From now on, $\bar{\mathbf{A}}^T$ will be denoted by $\mathbf{A}^*$, so $[\mathbf{A}^*]_{ij} = \overline{a}_{ji}$. For example,

$$\begin{pmatrix} 1-4\mathrm{i} & \mathrm{i} & 2 \\ 3 & 2+\mathrm{i} & 0 \end{pmatrix}^* = \begin{pmatrix} 1+4\mathrm{i} & 3 \\ -\mathrm{i} & 2-\mathrm{i} \\ 2 & 0 \end{pmatrix}.$$

$(\mathbf{A}^*)^* = \mathbf{A}$ for all matrices, and $\mathbf{A}^* = \mathbf{A}^T$ whenever $\mathbf{A}$ contains only real entries. Sometimes the matrix $\mathbf{A}^*$ is called the ***adjoint*** of $\mathbf{A}$.

The transpose (and conjugate transpose) operation is easily combined with matrix addition and scalar multiplication. The basic rules are given below.

Properties of the Transpose

If $\mathbf{A}$ and $\mathbf{B}$ are two matrices of the same shape, and if α is a scalar, then each of the following statements is true.

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \quad \text{and} \quad (\mathbf{A}+\mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*. \tag{3.2.1}$$

$$(\alpha\mathbf{A})^T = \alpha\mathbf{A}^T \quad \text{and} \quad (\alpha\mathbf{A})^* = \overline{\alpha}\mathbf{A}^*. \tag{3.2.2}$$

Proof.[17] We will prove that (3.2.1) and (3.2.2) hold for the transpose operation. The proofs of the statements involving conjugate transposes are similar and are left as exercises. For each i and j, it is true that

$$[(\mathbf{A}+\mathbf{B})^T]_{ij} = [\mathbf{A}+\mathbf{B}]_{ji} = [\mathbf{A}]_{ji} + [\mathbf{B}]_{ji} = [\mathbf{A}^T]_{ij} + [\mathbf{B}^T]_{ij} = [\mathbf{A}^T + \mathbf{B}^T]_{ij}.$$

[17] Computers can outperform people in many respects in that they do arithmetic much faster and more accurately than we can, and they are now rather adept at symbolic computation and mechanical manipulation of formulas. But computers can't do mathematics—people still hold the monopoly. Mathematics emanates from the uniquely human capacity to reason abstractly in a creative and logical manner, and learning mathematics goes hand-in-hand with learning how to reason abstractly and create logical arguments. This is true regardless of whether your orientation is applied or theoretical. For this reason, formal proofs will appear more frequently as the text evolves, and it is expected that your level of comprehension as well as your ability to create proofs will grow as you proceed.

This proves that corresponding entries in $(\mathbf{A}+\mathbf{B})^T$ and $\mathbf{A}^T+\mathbf{B}^T$ are equal, so it must be the case that $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T+\mathbf{B}^T$. Similarly, for each i and j,

$$[(\alpha\mathbf{A})^T]_{ij} = [\alpha\mathbf{A}]_{ji} = \alpha[\mathbf{A}]_{ji} = \alpha[\mathbf{A}^T]_{ij} \quad\Longrightarrow\quad (\alpha\mathbf{A})^T = \alpha\mathbf{A}^T. \quad \blacksquare$$

Sometimes transposition doesn't change anything. For example, if

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

This is because the entries in $\mathbf{A}$ are symmetrically located about the ***main diagonal***—the line from the upper-left-hand corner to the lower-right-hand corner. Matrices of the form $\mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$ are called ***diagonal matrices***, and they are clearly symmetric in the sense that $\mathbf{D} = \mathbf{D}^T$. This is one of several kinds of symmetries described below.

Symmetries

Let $\mathbf{A} = [a_{ij}]$ be a square matrix.

- $\mathbf{A}$ is said to be a ***symmetric matrix*** whenever $\mathbf{A} = \mathbf{A}^T$, i.e., whenever $a_{ij} = a_{ji}$.
- $\mathbf{A}$ is said to be a ***skew-symmetric matrix*** whenever $\mathbf{A} = -\mathbf{A}^T$, i.e., whenever $a_{ij} = -a_{ji}$.
- $\mathbf{A}$ is said to be a ***hermitian matrix*** whenever $\mathbf{A} = \mathbf{A}^*$, i.e., whenever $a_{ij} = \overline{a}_{ji}$. This is the complex analog of symmetry.
- $\mathbf{A}$ is said to be a ***skew-hermitian matrix*** when $\mathbf{A} = -\mathbf{A}^*$, i.e., whenever $a_{ij} = -\overline{a}_{ji}$. This is the complex analog of skew symmetry.

For example, consider

$$\mathbf{A} = \begin{pmatrix} 1 & 2+4\mathrm{i} & 1-3\mathrm{i} \\ 2-4\mathrm{i} & 3 & 8+6\mathrm{i} \\ 1+3\mathrm{i} & 8-6\mathrm{i} & 5 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 2+4\mathrm{i} & 1-3\mathrm{i} \\ 2+4\mathrm{i} & 3 & 8+6\mathrm{i} \\ 1-3\mathrm{i} & 8+6\mathrm{i} & 5 \end{pmatrix}.$$

Can you see that $\mathbf{A}$ is hermitian but not symmetric, while $\mathbf{B}$ is symmetric but not hermitian?

Nature abounds with symmetry, and very often physical symmetry manifests itself as a symmetric matrix in a mathematical model. The following example is an illustration of this principle.

Example 3.2.1

Consider two springs that are connected as shown in Figure 3.2.1.

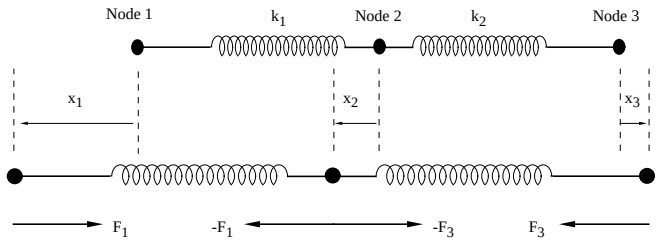


FIGURE 3.2.1

The springs at the top represent the "no tension" position in which no force is being exerted on any of the nodes. Suppose that the springs are stretched or compressed so that the nodes are displaced as indicated in the lower portion of Figure 3.2.1. Stretching or compressing the springs creates a force on each node according to Hooke's law [18] that says that the force exerted by a spring is $F = kx$, where x is the distance the spring is stretched or compressed and where k is a *stiffness constant* inherent to the spring. Suppose our springs have stiffness constants k_1 and k_2, and let F_i be the force on node i when the springs are stretched or compressed. Let's agree that a displacement to the left is positive, while a displacement to the right is negative, and consider a force directed to the right to be positive while one directed to the left is negative. If node 1 is displaced x_1 units, and if node 2 is displaced x_2 units, then the left-hand spring is stretched (or compressed) by a total amount of $x_1 - x_2$ units, so the force on node 1 is

$$F_1 = k_1(x_1 - x_2).$$

Similarly, if node 2 is displaced x_2 units, and if node 3 is displaced x_3 units, then the right-hand spring is stretched by a total amount of $x_2 - x_3$ units, so the force on node 3 is

$$F_3 = -k_2(x_2 - x_3).$$

The minus sign indicates the force is directed to the left. The force on the left-hand side of node 2 is the opposite of the force on node 1, while the force on the right-hand side of node 2 must be the opposite of the force on node 3. That is,

$$F_2 = -F_1 - F_3.$$

[18] Hooke's law is named for Robert Hooke (1635–1703), an English physicist, but it was generally known to several people (including Newton) before Hooke's 1678 claim to it was made. Hooke was a creative person who is credited with several inventions, including the wheel barometer, but he was reputed to be a man of "terrible character." This characteristic virtually destroyed his scientific career as well as his personal life. It is said that he lacked mathematical sophistication and that he left much of his work in incomplete form, but he bitterly resented people who built on his ideas by expressing them in terms of elegant mathematical formulations.

Organize the above three equations as a linear system:

$$\begin{aligned} k_1x_1 \qquad\qquad - k_1x_2 \qquad\qquad &= F_1, \\ -k_1x_1 + (k_1+k_2)x_2 - k_2x_3 &= F_2, \\ -k_2x_2 + k_2x_3 &= F_3, \end{aligned}$$

and observe that the coefficient matrix, called the ***stiffness matrix,***

$$\mathbf{K} = \begin{pmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1+k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{pmatrix},$$

is a *symmetric* matrix. The point of this example is that symmetry in the physical problem translates to symmetry in the mathematics by way of the symmetric matrix $\mathbf{K}$. When the two springs are identical (i.e., when $k_1 = k_2 = k$), even more symmetry is present, and in this case

$$\mathbf{K} = k\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Exercises for section 3.2

3.2.1. Determine the unknown quantities in the following expressions.

(a) $3\mathbf{X} = \begin{pmatrix} 0 & 3 \\ 6 & 9 \end{pmatrix}$. (b) $2\begin{pmatrix} x+2 & y+3 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ y & z \end{pmatrix}^T$.

3.2.2. Identify each of the following as symmetric, skew symmetric, or neither.

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

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

3.2.3. Construct an example of a 3×3 matrix $\mathbf{A}$ that satisfies the following conditions.

(a) $\mathbf{A}$ is both symmetric and skew symmetric.
(b) $\mathbf{A}$ is both hermitian and symmetric.
(c) $\mathbf{A}$ is skew hermitian.

3.2.4. Explain why the set of all $n \times n$ symmetric matrices is closed under matrix addition. That is, explain why the sum of two $n \times n$ symmetric matrices is again an $n \times n$ symmetric matrix. Is the set of all $n \times n$ skew-symmetric matrices closed under matrix addition?

3.2.5. Prove that each of the following statements is true.

(a) If $\mathbf{A} = [a_{ij}]$ is skew symmetric, then $a_{jj} = 0$ for each j.

(b) If $\mathbf{A} = [a_{ij}]$ is skew hermitian, then each a_{jj} is a pure imaginary number—i.e., a multiple of the imaginary unit i.

(c) If $\mathbf{A}$ is real and symmetric, then $\mathbf{B} = \mathrm{i}\mathbf{A}$ is skew hermitian.

3.2.6. Let $\mathbf{A}$ be any square matrix.

(a) Show that $\mathbf{A}+\mathbf{A}^T$ is symmetric and $\mathbf{A}-\mathbf{A}^T$ is skew symmetric.

(b) Prove that there is one and only one way to write $\mathbf{A}$ as the sum of a symmetric matrix and a skew-symmetric matrix.

3.2.7. If $\mathbf{A}$ and $\mathbf{B}$ are two matrices of the same shape, prove that each of the following statements is true.

(a) $(\mathbf{A}+\mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*$.

(b) $(\alpha\mathbf{A})^* = \overline{\alpha}\mathbf{A}^*$.

3.2.8. Using the conventions given in Example 3.2.1, determine the stiffness matrix for a system of n identical springs, with stiffness constant k, connected in a line similar to that shown in Figure 3.2.1.

3.3 LINEARITY

The concept of linearity is the underlying theme of our subject. In elementary mathematics the term "linear function" refers to straight lines, but in higher mathematics linearity means something much more general. Recall that a function f is simply a rule for associating points in one set $\mathcal{D}$—called the ***domain*** of f—to points in another set $\mathcal{R}$—the ***range*** of f. A *linear* function is a particular type of function that is characterized by the following two properties.

Linear Functions

Suppose that $\mathcal{D}$ and $\mathcal{R}$ are sets that possess an addition operation as well as a scalar multiplication operation—i.e., a multiplication between scalars and set members. A function f that maps points in $\mathcal{D}$ to points in $\mathcal{R}$ is said to be a ***linear function*** whenever f satisfies the conditions that

$$f(x+y) = f(x) + f(y) \tag{3.3.1}$$

and

$$f(\alpha x) = \alpha f(x) \tag{3.3.2}$$

for every x and y in $\mathcal{D}$ and for all scalars α. These two conditions may be combined by saying that f is a linear function whenever

$$f(\alpha x + y) = \alpha f(x) + f(y) \tag{3.3.3}$$

for all scalars α and for all $x, y \in \mathcal{D}$.

One of the simplest linear functions is $f(x) = \alpha x$, whose graph in $\Re^2$ is a straight line through the origin. You should convince yourself that f is indeed a linear function according to the above definition. However, $f(x) = \alpha x + \beta$ does not qualify for the title "linear function"—it is a linear function that has been translated by a constant β. Translations of linear functions are referred to as ***affine functions.*** Virtually all information concerning affine functions can be derived from an understanding of linear functions, and consequently we will focus only on issues of linearity.

In $\Re^3$, the surface described by a function of the form

$$f(x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2$$

is a plane through the origin, and it is easy to verify that f is a linear function. For $\beta \neq 0$, the graph of $f(x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \beta$ is a plane *not* passing through the origin, and f is no longer a linear function—it is an affine function.

In $\Re^2$ and $\Re^3$, the graphs of linear functions are lines and planes through the origin, and there seems to be a pattern forming. Although we cannot visualize higher dimensions with our eyes, it seems reasonable to suggest that a general linear function of the form

$$f(x_1, x_2, \ldots, x_n) = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

somehow represents a "linear" or "flat" surface passing through the origin $\mathbf{0} = (0, 0, \ldots, 0)$ in $\Re^{n+1}$. One of the goals of the next chapter is to learn how to better interpret and understand this statement.

Linearity is encountered at every turn. For example, the familiar operations of differentiation and integration may be viewed as linear functions. Since

$$\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx} \quad \text{and} \quad \frac{d(\alpha f)}{dx} = \alpha \frac{df}{dx},$$

the differentiation operator $D_x(f) = df/dx$ is linear. Similarly,

$$\int (f+g)dx = \int f dx + \int g dx \quad \text{and} \quad \int \alpha f dx = \alpha \int f dx$$

means that the integration operator $I(f) = \int f dx$ is linear.

There are several important matrix functions that are linear. For example, the transposition function $f(\mathbf{X}_{m\times n}) = \mathbf{X}^T$ is linear because

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \quad \text{and} \quad (\alpha \mathbf{A})^T = \alpha \mathbf{A}^T$$

(recall (3.2.1) and (3.2.2)). Another matrix function that is linear is the *trace* function presented below.

Example 3.3.1

The ***trace*** of an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$ is defined to be the sum of the entries lying on the main diagonal of $\mathbf{A}$. That is,

$$trace\,(\mathbf{A}) = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^{n} a_{ii}.$$

Problem: Show that $f(\mathbf{X}_{n\times n}) = trace\,(\mathbf{X})$ is a linear function.

Solution: Let's be efficient by showing that (3.3.3) holds. Let $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{ij}]$, and write

$$\begin{aligned} f(\alpha\mathbf{A} + \mathbf{B}) &= trace\,(\alpha\mathbf{A} + \mathbf{B}) = \sum_{i=1}^{n} [\alpha\mathbf{A} + \mathbf{B}]_{ii} = \sum_{i=1}^{n} (\alpha a_{ii} + b_{ii}) \\ &= \sum_{i=1}^{n} \alpha a_{ii} + \sum_{i=1}^{n} b_{ii} = \alpha \sum_{i=1}^{n} a_{ii} + \sum_{i=1}^{n} b_{ii} = \alpha\, trace\,(\mathbf{A}) + trace\,(\mathbf{B}) \\ &= \alpha f(\mathbf{A}) + f(\mathbf{B}). \end{aligned}$$

Example 3.3.2

Consider a linear system

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & u_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & u_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & u_m, \end{array}$$

to be a function $\mathbf{u} = f(\mathbf{x})$ that maps $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \Re^n$ to $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} \in \Re^m$.

Problem: Show that $\mathbf{u} = f(\mathbf{x})$ is linear.

Solution: Let $\mathbf{A} = [a_{ij}]$ be the matrix of coefficients, and write

$$f(\alpha\mathbf{x} + \mathbf{y}) = f\begin{pmatrix} \alpha x_1 + y_1 \\ \alpha x_2 + y_2 \\ \vdots \\ \alpha x_n + y_n \end{pmatrix} = \sum_{j=1}^{n} (\alpha x_j + y_j)\mathbf{A}_{*j} = \sum_{j=1}^{n} (\alpha x_j \mathbf{A}_{*j} + y_j \mathbf{A}_{*j})$$

$$= \sum_{j=1}^{n} \alpha x_j \mathbf{A}_{*j} + \sum_{j=1}^{n} y_j \mathbf{A}_{*j} = \alpha \sum_{j=1}^{n} x_j \mathbf{A}_{*j} + \sum_{j=1}^{n} y_j \mathbf{A}_{*j}$$

$$= \alpha f(\mathbf{x}) + f(\mathbf{y}).$$

According to (3.3.3), the function f is linear.

The following terminology will be used from now on.

Linear Combinations

For scalars α_j and matrices $\mathbf{X}_j$, the expression

$$\alpha_1\mathbf{X}_1 + \alpha_2\mathbf{X}_2 + \cdots + \alpha_n\mathbf{X}_n = \sum_{j=1}^{n} \alpha_j \mathbf{X}_j$$

is called a ***linear combination*** of the $\mathbf{X}_j$'s.

Exercises for section 3.3

3.3.1. Each of the following is a function from $\Re^2$ into $\Re^2$. Determine which are linear functions.

(a) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 1+y \end{pmatrix}$. (b) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x \end{pmatrix}$.

(c) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ xy \end{pmatrix}$. (d) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^2 \\ y^2 \end{pmatrix}$.

(e) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ \sin y \end{pmatrix}$. (f) $f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y \\ x-y \end{pmatrix}$.

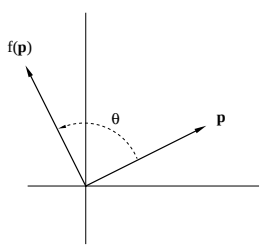
3.3.2. For $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$, and for constants ξ_i, verify that

$$f(\mathbf{x}) = \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_n x_n$$

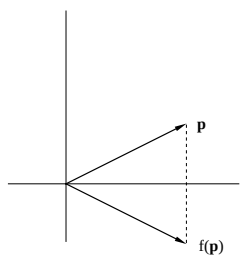
is a linear function.

3.3.3. Give examples of at least two different physical principles or laws that can be characterized as being linear phenomena.

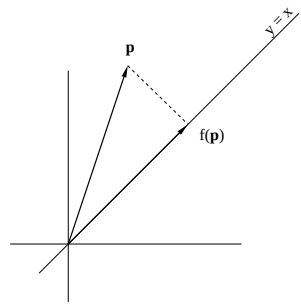
3.3.4. Determine which of the following three transformations in $\Re^2$ are linear.



Rotate counterclockwise through an angle θ.



Reflect about the x-axis.



Project onto the line $y = x$.

3.4 WHY DO IT THIS WAY

If you were given the task of formulating a definition for composing two matrices $\mathbf{A}$ and $\mathbf{B}$ in some sort of "natural" multiplicative fashion, your first attempt would probably be to compose $\mathbf{A}$ and $\mathbf{B}$ by multiplying corresponding entries—much the same way matrix addition is defined. Asked then to defend the usefulness of such a definition, you might be hard pressed to provide a truly satisfying response. Unless a person is in the right frame of mind, the issue of deciding how to best define matrix multiplication is not at all transparent, especially if it is insisted that the definition be both "natural" and "useful." The world had to wait for Arthur Cayley to come to this proper frame of mind.

As mentioned in §3.1, matrix algebra appeared late in the game. Manipulation on arrays and the theory of determinants existed long before Cayley and his theory of matrices. Perhaps this can be attributed to the fact that the "correct" way to multiply two matrices eluded discovery for such a long time.

Around 1855, Cayley became interested in composing linear functions.[19] In particular, he was investigating linear functions of the type discussed in Example 3.3.2. Typical examples of two such functions are

$$f(\mathbf{x}) = f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{pmatrix} \quad \text{and} \quad g(\mathbf{x}) = g\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} Ax_1 + Bx_2 \\ Cx_1 + Dx_2 \end{pmatrix}.$$

Consider, as Cayley did, composing f and g to create another linear function

$$h(\mathbf{x}) = f\Big(g(\mathbf{x})\Big) = f\begin{pmatrix} Ax_1 + Bx_2 \\ Cx_1 + Dx_2 \end{pmatrix} = \begin{pmatrix} (aA + bC)x_1 + (aB + bD)x_2 \\ (cA + dC)x_1 + (cB + dD)x_2 \end{pmatrix}.$$

It was Cayley's idea to use matrices of coefficients to represent these linear functions. That is, f, g, and h are represented by

$$\mathbf{F} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \end{pmatrix}.$$

After making this association, it was only natural for Cayley to call $\mathbf{H}$ the *composition* (or *product*) of $\mathbf{F}$ and $\mathbf{G}$, and to write

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \end{pmatrix}. \tag{3.4.1}$$

In other words, the product of two matrices represents the composition of the two associated linear functions. By means of this observation, Cayley brought to life the subjects of matrix analysis and linear algebra.

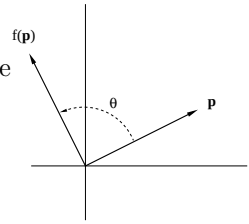
---

[19] Cayley was not the first to compose linear functions. In fact, Gauss used these compositions as early as 1801, but not in the form of an array of coefficients. Cayley was the first to make the connection between composition of linear functions and the composition of the associated matrices. Cayley's work from 1855 to 1857 is regarded as being the birth of our subject.

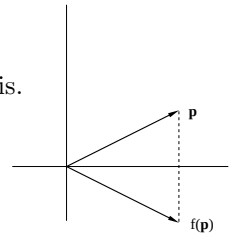
Exercises for section 3.4

Each problem in this section concerns the following three linear transformations in $\Re^2$.

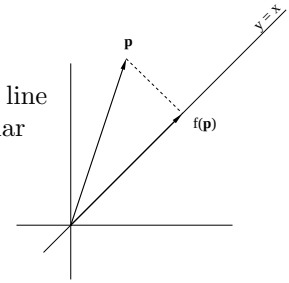
Rotation: Rotate points counterclockwise through an angle θ.



Reflection: Reflect points about the x-axis.



Projection: Project points onto the line $y = x$ in a perpendicular manner.



3.4.1. Determine the matrix associated with each of these linear functions. That is, determine the a_{ij}'s such that

$$f(\mathbf{p}) = f\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}.$$

3.4.2. By using matrix multiplication, determine the linear function obtained by performing a rotation followed by a reflection.

3.4.3. By using matrix multiplication, determine the linear function obtained by first performing a reflection, then a rotation, and finally a projection.

3.5 MATRIX MULTIPLICATION

The purpose of this section is to further develop the concept of matrix multiplication as introduced in the previous section. In order to do this, it is helpful to begin by composing a single row with a single column. If

$$\mathbf{R} = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix},$$

the ***standard inner product*** of $\mathbf{R}$ with $\mathbf{C}$ is defined to be the scalar

$$\mathbf{RC} = r_1c_1 + r_2c_2 + \cdots + r_nc_n = \sum_{i=1}^{n} r_ic_i.$$

For example,

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

Recall from (3.4.1) that the product of two 2×2 matrices

$$\mathbf{F} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \mathbf{G} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

was defined naturally by writing

$$\mathbf{FG} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \end{pmatrix} = \mathbf{H}.$$

Notice that the (i, j)-entry in the product $\mathbf{H}$ can be described as the inner product of the ith row of $\mathbf{F}$ with the jth column in $\mathbf{G}$. That is,

$$h_{11} = \mathbf{F}_{1*}\mathbf{G}_{*1} = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} A \\ C \end{pmatrix}, \qquad h_{12} = \mathbf{F}_{1*}\mathbf{G}_{*2} = \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} B \\ D \end{pmatrix},$$

$$h_{21} = \mathbf{F}_{2*}\mathbf{G}_{*1} = \begin{pmatrix} c & d \end{pmatrix} \begin{pmatrix} A \\ C \end{pmatrix}, \qquad h_{22} = \mathbf{F}_{2*}\mathbf{G}_{*2} = \begin{pmatrix} c & d \end{pmatrix} \begin{pmatrix} B \\ D \end{pmatrix}.$$

This is exactly the way that the general definition of matrix multiplication is formulated.

Matrix Multiplication

- Matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be ***conformable*** for multiplication in the order $\mathbf{AB}$ whenever $\mathbf{A}$ has exactly as many columns as $\mathbf{B}$ has rows—i.e., $\mathbf{A}$ is $m \times p$ and $\mathbf{B}$ is $p \times n$.

- For conformable matrices $\mathbf{A}_{m\times p} = [a_{ij}]$ and $\mathbf{B}_{p\times n} = [b_{ij}]$, the ***matrix product*** $\mathbf{AB}$ is defined to be the $m \times n$ matrix whose (i,j)-entry is the inner product of the ith row of $\mathbf{A}$ with the jth column in $\mathbf{B}$. That is,

$$[\mathbf{AB}]_{ij} = \mathbf{A}_{i*}\mathbf{B}_{*j} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ip}b_{pj} = \sum_{k=1}^{p} a_{ik}b_{kj}.$$

- In case $\mathbf{A}$ and $\mathbf{B}$ fail to be conformable—i.e., $\mathbf{A}$ is $m \times p$ and $\mathbf{B}$ is $q \times n$ with $p \neq q$—then no product $\mathbf{AB}$ is defined.

For example, if

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}_{2\times 3} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{pmatrix}_{3\times 4}$$

inside ones match

shape of the product

then the product $\mathbf{AB}$ exists and has shape 2×4. Consider a typical entry of this product, say, the (2,3)-entry. The definition says $[\mathbf{AB}]_{23}$ is obtained by forming the inner product of the second row of $\mathbf{A}$ with the third column of $\mathbf{B}$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ \boxed{a_{21} \quad a_{22} \quad a_{23}} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \boxed{b_{13}} & b_{14} \\ b_{21} & b_{22} & \boxed{b_{23}} & b_{24} \\ b_{31} & b_{32} & \boxed{b_{33}} & b_{34} \end{pmatrix},$$

so

$$[\mathbf{AB}]_{23} = \mathbf{A}_{2*}\mathbf{B}_{*3} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} = \sum_{k=1}^{3} a_{2k}b_{k3}.$$

For example,

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -4 \\ -3 & 0 & 5 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 1 & 3 & -3 & 2 \\ 2 & 5 & -1 & 8 \\ -1 & 2 & 0 & 2 \end{pmatrix} \Longrightarrow \mathbf{AB} = \begin{pmatrix} 8 & 3 & -7 & 4 \\ -8 & 1 & 9 & 4 \end{pmatrix}.$$

Notice that in spite of the fact that the product $\mathbf{AB}$ exists, the product $\mathbf{BA}$ is not defined—matrix $\mathbf{B}$ is 3×4 and $\mathbf{A}$ is 2×3, and the inside dimensions don't match in this order. Even when the products $\mathbf{AB}$ and $\mathbf{BA}$ each exist and have the same shape, they need not be equal. For example,

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \implies \mathbf{AB} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \ \mathbf{BA} = \begin{pmatrix} 2 & -2 \\ 2 & -2 \end{pmatrix}. \tag{3.5.1}$$

This disturbing feature is a primary difference between scalar and matrix algebra.

Matrix Multiplication Is Not Commutative

Matrix multiplication is a noncommutative operation—i.e., it is possible for $\mathbf{AB} \neq \mathbf{BA}$, even when both products exist and have the same shape.

There are other major differences between multiplication of matrices and multiplication of scalars. For scalars,

$$\alpha\beta = 0 \quad \text{implies} \quad \alpha = 0 \quad \text{or} \quad \beta = 0. \tag{3.5.2}$$

However, the analogous statement for matrices does not hold—the matrices given in (3.5.1) show that it is possible for $\mathbf{AB} = \mathbf{0}$ with $\mathbf{A} \neq \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$. Related to this issue is a rule sometimes known as the ***cancellation law.*** For scalars, this law says that

$$\alpha\beta = \alpha\gamma \quad \text{and} \quad \alpha \neq 0 \quad \text{implies} \quad \beta = \gamma. \tag{3.5.3}$$

This is true because we invoke (3.5.2) to deduce that $\alpha(\beta - \gamma) = 0$ implies $\beta - \gamma = 0$. Since (3.5.2) does not hold for matrices, we cannot expect (3.5.3) to hold for matrices.

Example 3.5.1

The cancellation law (3.5.3) fails for matrix multiplication. If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$

then

$$\mathbf{AB} = \begin{pmatrix} 4 & 4 \\ 4 & 4 \end{pmatrix} = \mathbf{AC} \quad \text{but} \quad \mathbf{B} \neq \mathbf{C}$$

in spite of the fact that $\mathbf{A} \neq \mathbf{0}$.

There are various ways to express the individual rows and columns of a matrix product. For example, the ith row of $\mathbf{AB}$ is

$$[\mathbf{AB}]_{i*} = \left[\mathbf{A}_{i*}\mathbf{B}_{*1} \,|\, \mathbf{A}_{i*}\mathbf{B}_{*2} \,|\, \cdots \,|\, \mathbf{A}_{i*}\mathbf{B}_{*n}\right] = \mathbf{A}_{i*}\mathbf{B}$$

$$= \begin{pmatrix} a_{i1} & a_{i2} & \cdots & a_{ip} \end{pmatrix} \begin{pmatrix} \mathbf{B}_{1*} \\ \mathbf{B}_{2*} \\ \vdots \\ \mathbf{B}_{p*} \end{pmatrix} = a_{i1}\mathbf{B}_{1*} + a_{i2}\mathbf{B}_{2*} + \cdots + a_{ip}\mathbf{B}_{p*}.$$

As shown below, there are similar representations for the individual columns.

Rows and Columns of a Product

Suppose that $\mathbf{A} = [a_{ij}]$ is $m \times p$ and $\mathbf{B} = [b_{ij}]$ is $p \times n$.

- $[\mathbf{AB}]_{i*} = \mathbf{A}_{i*}\mathbf{B}$ $\left[(\, i^{th} \text{ row of } \mathbf{AB}\,) = (\, i^{th} \text{ row of } \mathbf{A}\,) \times \mathbf{B}\right]$. (3.5.4)
- $[\mathbf{AB}]_{*j} = \mathbf{AB}_{*j}$ $\left[(\, j^{th} \text{ col of } \mathbf{AB}\,) = \mathbf{A} \times (\, j^{th} \text{ col of } \mathbf{B}\,)\right]$. (3.5.5)
- $[\mathbf{AB}]_{i*} = a_{i1}\mathbf{B}_{1*} + a_{i2}\mathbf{B}_{2*} + \cdots + a_{ip}\mathbf{B}_{p*} = \sum_{k=1}^{p} a_{ik}\mathbf{B}_{k*}$. (3.5.6)
- $[\mathbf{AB}]_{*j} = \mathbf{A}_{*1}b_{1j} + \mathbf{A}_{*2}b_{2j} + \cdots + \mathbf{A}_{*p}b_{pj} = \sum_{k=1}^{p} \mathbf{A}_{*k}b_{kj}$. (3.5.7)

These last two equations show that rows of $\mathbf{AB}$ are combinations of rows of $\mathbf{B}$, while columns of $\mathbf{AB}$ are combinations of columns of $\mathbf{A}$.

For example, if $\mathbf{A} = \begin{pmatrix} 1 & -2 & 0 \\ 3 & -4 & 5 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & -5 & 1 \\ 2 & -7 & 2 \\ 1 & -2 & 0 \end{pmatrix}$, then the second row of $\mathbf{AB}$ is

$$[\mathbf{AB}]_{2*} = \mathbf{A}_{2*}\mathbf{B} = \begin{pmatrix} 3 & -4 & 5 \end{pmatrix} \begin{pmatrix} 3 & -5 & 1 \\ 2 & -7 & 2 \\ 1 & -2 & 0 \end{pmatrix} = \begin{pmatrix} 6 & 3 & -5 \end{pmatrix},$$

and the second column of $\mathbf{AB}$ is

$$[\mathbf{AB}]_{*2} = \mathbf{AB}_{*2} = \begin{pmatrix} 1 & -2 & 0 \\ 3 & -4 & 5 \end{pmatrix} \begin{pmatrix} -5 \\ -7 \\ -2 \end{pmatrix} = \begin{pmatrix} 9 \\ 3 \end{pmatrix}.$$

This example makes the point that it is wasted effort to compute the entire product if only one row or column is called for. Although it's not necessary to compute the complete product, you may wish to verify that

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

Matrix multiplication provides a convenient representation for a linear system of equations. For example, the 3×4 system

$$\begin{aligned} 2x_1 + 3x_2 + 4x_3 + 8x_4 &= 7, \\ 3x_1 + 5x_2 + 6x_3 + 2x_4 &= 6, \\ 4x_1 + 2x_2 + 4x_3 + 9x_4 &= 4, \end{aligned}$$

can be written as $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A}_{3\times 4} = \begin{pmatrix} 2 & 3 & 4 & 8 \\ 3 & 5 & 6 & 2 \\ 4 & 2 & 4 & 9 \end{pmatrix}, \quad \mathbf{x}_{4\times 1} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad \text{and} \quad \mathbf{b}_{3\times 1} = \begin{pmatrix} 7 \\ 6 \\ 4 \end{pmatrix}.$$

And this example generalizes to become the following statement.

Linear Systems

Every linear system of m equations in n unknowns

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1, \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2, \\ & & & & \vdots & & & & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m, \end{array}$$

can be written as a single matrix equation $\mathbf{Ax} = \mathbf{b}$ in which

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Conversely, every matrix equation of the form $\mathbf{A}_{m\times n}\mathbf{x}_{n\times 1} = \mathbf{b}_{m\times 1}$ represents a system of m linear equations in n unknowns.

The numerical solution of a linear system was presented earlier in the text without the aid of matrix multiplication because the operation of matrix multiplication is not an integral part of the arithmetical process used to extract a solution by means of Gaussian elimination. Viewing a linear system as a single matrix equation $\mathbf{Ax} = \mathbf{b}$ is more of a notational convenience that can be used to uncover theoretical properties and to prove general theorems concerning linear systems.

For example, a very concise proof of the fact (2.3.5) stating that a system of equations $\mathbf{A}_{m\times n}\mathbf{x}_{n\times 1} = \mathbf{b}_{m\times 1}$ is consistent if and only if $\mathbf{b}$ is a linear combination of the columns in $\mathbf{A}$ is obtained by noting that the system is consistent if and only if there exists a column $\mathbf{s}$ that satisfies

$$\mathbf{b} = \mathbf{A}\mathbf{s} = \begin{pmatrix} \mathbf{A}_{*1} & \mathbf{A}_{*2} & \cdots & \mathbf{A}_{*n} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = \mathbf{A}_{*1}s_1 + \mathbf{A}_{*2}s_2 + \cdots + \mathbf{A}_{*n}s_n.$$

The following example illustrates a common situation in which matrix multiplication arises naturally.

Example 3.5.2

An airline serves five cities, say, A, B, C, D, and H, in which H is the "hub city." The various routes between the cities are indicated in Figure 3.5.1.

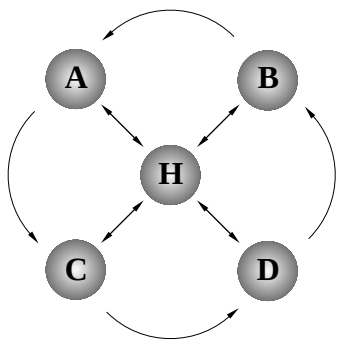


FIGURE 3.5.1

Suppose you wish to travel from city A to city B so that at least two connecting flights are required to make the trip. Flights $(A \to H)$ and $(H \to B)$ provide the minimal number of connections. However, if space on either of these two flights is not available, you will have to make at least three flights. Several questions arise. How many routes from city A to city B require *exactly* three connecting flights? How many routes require *no more than* four flights—and so forth? Since this particular network is small, these questions can be answered by "eyeballing" the diagram, but the "eyeball method" won't get you very far with the large networks that occur in more practical situations. Let's see how matrix algebra can be applied. Begin by creating a ***connectivity matrix*** $\mathbf{C} = [c_{ij}]$ (also known as an ***adjacency matrix***) in which

$$c_{ij} = \begin{cases} 1 & \text{if there is a flight from city } i \text{ to city } j, \\ 0 & \text{otherwise.} \end{cases}$$

For the network depicted in Figure 3.5.1,

$$\mathbf{C} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ H \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & H \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

The matrix $\mathbf{C}$ together with its powers $\mathbf{C}^2, \mathbf{C}^3, \mathbf{C}^4, \ldots$ will provide all of the information needed to analyze the network. To see how, notice that since c_{ik} is the number of direct routes from city i to city k, and since c_{kj} is the number of direct routes from city k to city j, it follows that $c_{ik}c_{kj}$ must be the number of 2-flight routes from city i to city j that have a connection at city k. Consequently, the (i,j)-entry in the product $\mathbf{C}^2 = \mathbf{CC}$ is

$$[\mathbf{C}^2]_{ij} = \sum_{k=1}^{5} c_{ik}c_{kj} = \text{ the total number of 2-flight routes from city } i \text{ to city } j.$$

Similarly, the (i,j)-entry in the product $\mathbf{C}^3 = \mathbf{CCC}$ is

$$[\mathbf{C}^3]_{ij} = \sum_{k_1,k_2=1}^{5} c_{ik_1}c_{k_1k_2}c_{k_2j} = \text{ number of 3-flight routes from city } i \text{ to city } j,$$

and, in general,

$$[\mathbf{C}^n]_{ij} = \sum_{k_1,k_2,\cdots,k_{n-1}=1}^{5} c_{ik_1}c_{k_1k_2}\cdots c_{k_{n-2}k_{n-1}}c_{k_{n-1}j}$$

is the total number of n-flight routes from city i to city j. Therefore, the total number of routes from city i to city j that require *no more* than n flights must be given by

$$[\mathbf{C}]_{ij} + [\mathbf{C}^2]_{ij} + [\mathbf{C}^3]_{ij} + \cdots + [\mathbf{C}^n]_{ij} = [\mathbf{C} + \mathbf{C}^2 + \mathbf{C}^3 + \cdots + \mathbf{C}^n]_{ij}.$$

For our particular network,

$$\mathbf{C}^2 = \begin{pmatrix} 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 4 \end{pmatrix}, \quad \mathbf{C}^3 = \begin{pmatrix} 2 & 3 & 2 & 2 & 5 \\ 2 & 2 & 2 & 3 & 5 \\ 3 & 2 & 2 & 2 & 5 \\ 2 & 2 & 3 & 2 & 5 \\ 5 & 5 & 5 & 5 & 4 \end{pmatrix}, \quad \mathbf{C}^4 = \begin{pmatrix} 8 & 7 & 7 & 7 & 9 \\ 7 & 8 & 7 & 7 & 9 \\ 7 & 7 & 8 & 7 & 9 \\ 7 & 7 & 7 & 8 & 9 \\ 9 & 9 & 9 & 9 & 20 \end{pmatrix},$$

and

$$\mathbf{C}+\mathbf{C}^2+\mathbf{C}^3+\mathbf{C}^4=\begin{pmatrix}11&11&11&11&16\\11&11&11&11&16\\11&11&11&11&16\\11&11&11&11&16\\16&16&16&16&28\end{pmatrix}.$$

The fact that $[\mathbf{C}^3]_{12}=3$ means there are exactly 3 three-flight routes from city A to city B, and $[\mathbf{C}^4]_{12}=7$ means there are exactly 7 four-flight routes—try to identify them. Furthermore, $[\mathbf{C}+\mathbf{C}^2+\mathbf{C}^3+\mathbf{C}^4]_{12}=11$ means there are 11 routes from city A to city B that require no more than 4 flights.

Exercises for section 3.5

3.5.1. For $\mathbf{A}=\begin{pmatrix}1&-2&3\\0&-5&4\\4&-3&8\end{pmatrix}$, $\mathbf{B}=\begin{pmatrix}1&2\\0&4\\3&7\end{pmatrix}$, and $\mathbf{C}=\begin{pmatrix}1\\2\\3\end{pmatrix}$, compute the following products when possible.

(a) $\mathbf{AB}$, (b) $\mathbf{BA}$, (c) $\mathbf{CB}$, (d) $\mathbf{C}^T\mathbf{B}$, (e) $\mathbf{A}^2$, (f) $\mathbf{B}^2$, (g) $\mathbf{C}^T\mathbf{C}$, (h) $\mathbf{CC}^T$, (i) $\mathbf{BB}^T$, (j) $\mathbf{B}^T\mathbf{B}$, (k) $\mathbf{C}^T\mathbf{AC}$.

3.5.2. Consider the following system of equations:

$$\begin{aligned}2x_1+x_2+x_3&=3,\\4x_1+2x_3&=10,\\2x_1+2x_2&=-2.\end{aligned}$$

(a) Write the system as a matrix equation of the form $\mathbf{Ax}=\mathbf{b}$.
(b) Write the solution of the system as a column $\mathbf{s}$ and verify by matrix multiplication that $\mathbf{s}$ satisfies the equation $\mathbf{Ax}=\mathbf{b}$.
(c) Write $\mathbf{b}$ as a linear combination of the columns in $\mathbf{A}$.

3.5.3. Let $\mathbf{E}=\begin{pmatrix}1&0&0\\0&1&0\\3&0&1\end{pmatrix}$ and let $\mathbf{A}$ be an arbitrary 3×3 matrix.

(a) Describe the rows of $\mathbf{EA}$ in terms of the rows of $\mathbf{A}$.
(b) Describe the columns of $\mathbf{AE}$ in terms of the columns of $\mathbf{A}$.

3.5.4. Let $\mathbf{e}_j$ denote the jth ***unit column*** that contains a 1 in the jth position and zeros everywhere else. For a general matrix $\mathbf{A}_{n\times n}$, describe the following products. (a) $\mathbf{Ae}_j$ (b) $\mathbf{e}_i^T\mathbf{A}$ (c) $\mathbf{e}_i^T\mathbf{Ae}_j$

3.5.5. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are $m \times n$ matrices. If $\mathbf{Ax} = \mathbf{Bx}$ holds for all $n \times 1$ columns $\mathbf{x}$, prove that $\mathbf{A} = \mathbf{B}$. **Hint:** What happens when $\mathbf{x}$ is a unit column?

3.5.6. For $\mathbf{A} = \begin{pmatrix} 1/2 & \alpha \\ 0 & 1/2 \end{pmatrix}$, determine $\lim_{n\to\infty} \mathbf{A}^n$. **Hint:** Compute a few powers of $\mathbf{A}$ and try to deduce the general form of $\mathbf{A}^n$.

3.5.7. If $\mathbf{C}_{m\times 1}$ and $\mathbf{R}_{1\times n}$ are matrices consisting of a single column and a single row, respectively, then the matrix product $\mathbf{P}_{m\times n} = \mathbf{CR}$ is sometimes called the ***outer product*** of $\mathbf{C}$ with $\mathbf{R}$. For conformable matrices $\mathbf{A}$ and $\mathbf{B}$, explain how to write the product $\mathbf{AB}$ as a sum of outer products involving the columns of $\mathbf{A}$ and the rows of $\mathbf{B}$.

3.5.8. A square matrix $\mathbf{U} = [u_{ij}]$ is said to be ***upper triangular*** whenever $u_{ij} = 0$ for $i > j$—i.e., all entries below the main diagonal are 0.

(a) If $\mathbf{A}$ and $\mathbf{B}$ are two $n \times n$ upper-triangular matrices, explain why the product $\mathbf{AB}$ must also be upper triangular.

(b) If $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{n\times n}$ are upper triangular, what are the diagonal entries of $\mathbf{AB}$?

(c) $\mathbf{L}$ is ***lower triangular*** when $\ell_{ij} = 0$ for $i < j$. Is it true that the product of two $n \times n$ lower-triangular matrices is again lower triangular?

3.5.9. If $\mathbf{A} = [a_{ij}(t)]$ is a matrix whose entries are functions of a variable t, the ***derivative*** of $\mathbf{A}$ with respect to t is defined to be the matrix of derivatives. That is,

$$\frac{d\mathbf{A}}{dt} = \left[\frac{da_{ij}}{dt}\right].$$

Derive the *product rule for differentiation*

$$\frac{d(\mathbf{AB})}{dt} = \frac{d\mathbf{A}}{dt}\mathbf{B} + \mathbf{A}\frac{d\mathbf{B}}{dt}.$$

3.5.10. Let $\mathbf{C}_{n\times n}$ be the connectivity matrix associated with a network of n nodes such as that described in Example 3.5.2, and let $\mathbf{e}$ be the $n \times 1$ column of all 1's. In terms of the network, describe the entries in each of the following products.

(a) Interpret the product $\mathbf{Ce}$.

(b) Interpret the product $\mathbf{e}^T\mathbf{C}$.

3.5.11. Consider three tanks each containing V gallons of brine. The tanks are connected as shown in Figure 3.5.2, and all spigots are opened at once. As fresh water at the rate of r gal/sec is pumped into the top of the first tank, r gal/sec leaves from the bottom and flows into the next tank, and so on down the line—there are r gal/sec entering at the top and leaving through the bottom of each tank.

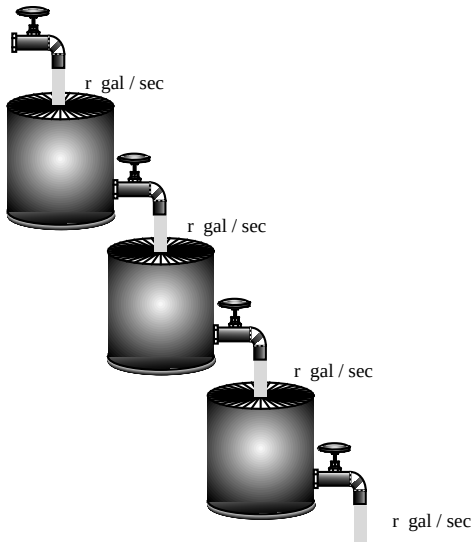


FIGURE 3.5.2

Let $x_i(t)$ denote the number of pounds of salt in tank i at time t, and let

$$\mathbf{x} = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} \quad \text{and} \quad \frac{d\mathbf{x}}{dt} = \begin{pmatrix} dx_1/dt \\ dx_2/dt \\ dx_3/dt \end{pmatrix}.$$

Assuming that complete mixing occurs in each tank on a continuous basis, show that

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}, \quad \text{where} \quad \mathbf{A} = \frac{r}{V}\begin{pmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Hint: Use the fact that

$$\frac{dx_i}{dt} = \text{ rate of change } = \frac{\text{lbs}}{\text{sec}} \text{ coming in} - \frac{\text{lbs}}{\text{sec}} \text{ going out.}$$

3.6 PROPERTIES OF MATRIX MULTIPLICATION

We saw in the previous section that there are some differences between scalar and matrix algebra—most notable is the fact that matrix multiplication is not commutative, and there is no cancellation law. But there are also some important similarities, and the purpose of this section is to look deeper into these issues.

Although we can adjust to not having the commutative property, the situation would be unbearable if the distributive and associative properties were not available. Fortunately, both of these properties hold for matrix multiplication.

Distributive and Associative Laws

For conformable matrices each of the following is true.

- $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$ (left-hand distributive law).
- $(\mathbf{D}+\mathbf{E})\mathbf{F} = \mathbf{DF}+\mathbf{EF}$ (right-hand distributive law).
- $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$ (associative law).

Proof. To prove the left-hand distributive property, demonstrate the corresponding entries in the matrices $\mathbf{A}(\mathbf{B}+\mathbf{C})$ and $\mathbf{AB}+\mathbf{AC}$ are equal. To this end, use the definition of matrix multiplication to write

$$\begin{aligned}
[\mathbf{A}(\mathbf{B}+\mathbf{C})]_{ij} &= \mathbf{A}_{i*}(\mathbf{B}+\mathbf{C})_{*j} = \sum_k [\mathbf{A}]_{ik}[\mathbf{B}+\mathbf{C}]_{kj} = \sum_k [\mathbf{A}]_{ik}\left([\mathbf{B}]_{kj}+[\mathbf{C}]_{kj}\right) \\
&= \sum_k \left([\mathbf{A}]_{ik}[\mathbf{B}]_{kj}+[\mathbf{A}]_{ik}[\mathbf{C}]_{kj}\right) = \sum_k [\mathbf{A}]_{ik}[\mathbf{B}]_{kj} + \sum_k [\mathbf{A}]_{ik}[\mathbf{C}]_{kj} \\
&= \mathbf{A}_{i*}\mathbf{B}_{*j} + \mathbf{A}_{i*}\mathbf{C}_{*j} = [\mathbf{AB}]_{ij} + [\mathbf{AC}]_{ij} \\
&= [\mathbf{AB}+\mathbf{AC}]_{ij}.
\end{aligned}$$

Since this is true for each i and j, it follows that $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$. The proof of the right-hand distributive property is similar and is omitted. To prove the associative law, suppose that $\mathbf{B}$ is $p \times q$ and $\mathbf{C}$ is $q \times n$, and recall from (3.5.7) that the jth column of $\mathbf{BC}$ is a linear combination of the columns in $\mathbf{B}$. That is,

$$[\mathbf{BC}]_{*j} = \mathbf{B}_{*1}c_{1j} + \mathbf{B}_{*2}c_{2j} + \cdots + \mathbf{B}_{*q}c_{qj} = \sum_{k=1}^{q} \mathbf{B}_{*k}c_{kj}.$$

Use this along with the left-hand distributive property to write

$$[\mathbf{A}(\mathbf{BC})]_{ij} = \mathbf{A}_{i*}[\mathbf{BC}]_{*j} = \mathbf{A}_{i*}\sum_{k=1}^{q}\mathbf{B}_{*k}c_{kj} = \sum_{k=1}^{q}\mathbf{A}_{i*}\mathbf{B}_{*k}c_{kj}$$
$$= \sum_{k=1}^{q}[\mathbf{AB}]_{ik}c_{kj} = [\mathbf{AB}]_{i*}\mathbf{C}_{*j} = [(\mathbf{AB})\mathbf{C}]_{ij}. \quad \blacksquare$$

Example 3.6.1

Linearity of Matrix Multiplication. Let $\mathbf{A}$ be an $m \times n$ matrix, and f be the function defined by matrix multiplication

$$f(\mathbf{X}_{n\times p}) = \mathbf{AX}.$$

The left-hand distributive property guarantees that f is a linear function because for all scalars α and for all $n \times p$ matrices $\mathbf{X}$ and $\mathbf{Y}$,

$$f(\alpha\mathbf{X}+\mathbf{Y}) = \mathbf{A}(\alpha\mathbf{X}+\mathbf{Y}) = \mathbf{A}(\alpha\mathbf{X}) + \mathbf{AY} = \alpha\mathbf{AX} + \mathbf{AY}$$
$$= \alpha f(\mathbf{X}) + f(\mathbf{Y}).$$

Of course, the linearity of matrix multiplication is no surprise because it was the consideration of linear functions that motivated the definition of the matrix product at the outset.

For scalars, the number 1 is the identity element for multiplication because it has the property that it reproduces whatever it is multiplied by. For matrices, there is an identity element with similar properties.

Identity Matrix

The $n \times n$ matrix with 1's on the main diagonal and 0's elsewhere

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

is called the ***identity matrix*** of order n. For every $m \times n$ matrix $\mathbf{A}$,

$$\mathbf{AI}_n = \mathbf{A} \quad \text{and} \quad \mathbf{I}_m\mathbf{A} = \mathbf{A}.$$

The subscript on $\mathbf{I}_n$ is neglected whenever the size is obvious from the context.

Proof. Notice that $\mathbf{I}_{*j}$ has a 1 in the jth position and 0's elsewhere. Recall from Exercise 3.5.4 that such columns were called ***unit columns,*** and they have the property that for any conformable matrix $\mathbf{A}$,

$$\mathbf{A}\mathbf{I}_{*j} = \mathbf{A}_{*j}.$$

Using this together with the fact that $[\mathbf{A}\mathbf{I}]_{*j} = \mathbf{A}\mathbf{I}_{*j}$ produces

$$\mathbf{A}\mathbf{I} = (\,\mathbf{A}\mathbf{I}_{*1} \quad \mathbf{A}\mathbf{I}_{*2} \quad \cdots \quad \mathbf{A}\mathbf{I}_{*n}\,) = (\,\mathbf{A}_{*1} \quad \mathbf{A}_{*2} \quad \cdots \quad \mathbf{A}_{*n}\,) = \mathbf{A}.$$

A similar argument holds when $\mathbf{I}$ appears on the left-hand side of $\mathbf{A}$. ∎

Analogous to scalar algebra, we define the 0th power of a square matrix to be the identity matrix of corresponding size. That is, if $\mathbf{A}$ is $n \times n$, then

$$\mathbf{A}^0 = \mathbf{I}_n.$$

Positive powers of $\mathbf{A}$ are also defined in the natural way. That is,

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

The associative law guarantees that it makes no difference how matrices are grouped for powering. For example, $\mathbf{A}\mathbf{A}^2$ is the same as $\mathbf{A}^2\mathbf{A}$, so that

$$\mathbf{A}^3 = \mathbf{A}\mathbf{A}\mathbf{A} = \mathbf{A}\mathbf{A}^2 = \mathbf{A}^2\mathbf{A}.$$

Also, the usual laws of exponents hold. For nonnegative integers r and s,

$$\mathbf{A}^r\mathbf{A}^s = \mathbf{A}^{r+s} \quad \text{and} \quad (\mathbf{A}^r)^s = \mathbf{A}^{rs}.$$

We are not yet in a position to define negative or fractional powers, and due to the lack of conformability, powers of nonsquare matrices are never defined.

Example 3.6.2

A Pitfall. For two $n \times n$ matrices, what is $(\mathbf{A}+\mathbf{B})^2$? **Be careful!** Because matrix multiplication is not commutative, the familiar formula from scalar algebra is not valid for matrices. The distributive properties must be used to write

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^2 &= \underbrace{(\mathbf{A}+\mathbf{B})}(\mathbf{A}+\mathbf{B}) = \underbrace{(\mathbf{A}+\mathbf{B})}\,\mathbf{A} + \underbrace{(\mathbf{A}+\mathbf{B})}\,\mathbf{B} \\ &= \mathbf{A}^2 + \mathbf{B}\mathbf{A} + \mathbf{A}\mathbf{B} + \mathbf{B}^2,\end{aligned}$$

and this is as far as you can go. The familiar form $\mathbf{A}^2+2\mathbf{A}\mathbf{B}+\mathbf{B}^2$ is obtained only in those rare cases where $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A}$. To evaluate $(\mathbf{A}+\mathbf{B})^k$, the distributive rules must be applied repeatedly, and the results are a bit more complicated—try it for $k = 3$.

Example 3.6.3

Suppose that the population migration between two geographical regions—say, the North and the South—is as follows. Each year, 50% of the population in the North migrates to the South, while only 25% of the population in the South moves to the North. This situation is depicted by drawing a transition diagram such as that shown in Figure 3.6.1.

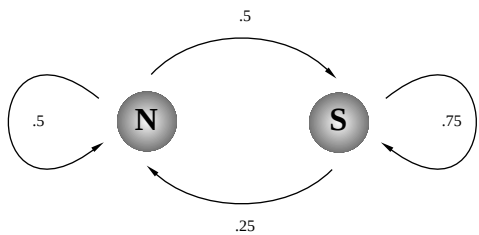


FIGURE 3.6.1

Problem: If this migration pattern continues, will the population in the North continually shrink until the entire population is eventually in the South, or will the population distribution somehow stabilize before the North is completely deserted?

Solution: Let n_k and s_k denote the respective proportions of the total population living in the North and South at the end of year k and assume $n_k + s_k = 1$. The migration pattern dictates that the fractions of the population in each region at the end of year $k+1$ are

$$\begin{aligned} n_{k+1} &= n_k(.5) + s_k(.25), \\ s_{k+1} &= n_k(.5) + s_k(.75). \end{aligned} \tag{3.6.1}$$

If $\mathbf{p}_k^T = (n_k, s_k)$ and $\mathbf{p}_{k+1}^T = (n_{k+1}, s_{k+1})$ denote the respective population distributions at the end of years k and $k+1$, and if

$$\mathbf{T} = \begin{array}{c} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{cc} \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} .5 & .5 \\ .25 & .75 \end{pmatrix} \end{array}$$

is the associated ***transition matrix,*** then (3.6.1) assumes the matrix form $\mathbf{p}_{k+1}^T = \mathbf{p}_k^T\mathbf{T}$. Inducting on $\mathbf{p}_1^T = \mathbf{p}_0^T\mathbf{T}$, $\mathbf{p}_2^T = \mathbf{p}_1^T\mathbf{T} = \mathbf{p}_0^T\mathbf{T}^2$, $\mathbf{p}_3^T = \mathbf{p}_2^T\mathbf{T} = \mathbf{p}_0^T\mathbf{T}^3$, etc., leads to

$$\mathbf{p}_k^T = \mathbf{p}_0^T\mathbf{T}^k. \tag{3.6.2}$$

Determining the long-run behavior involves evaluating $\lim_{k\to\infty} \mathbf{p}_k^T$, and it's clear from (3.6.2) that this boils down to analyzing $\lim_{k\to\infty} \mathbf{T}^k$. Later, in Example

7.3.5, a more sophisticated approach is discussed, but for now we will use the "brute force" method of successively powering $\mathbf{P}$ until a pattern emerges. The first several powers of $\mathbf{P}$ are shown below with three significant digits displayed.

$$\mathbf{P}^2 = \begin{pmatrix} .375 & .625 \\ .312 & .687 \end{pmatrix} \quad \mathbf{P}^3 = \begin{pmatrix} .344 & .656 \\ .328 & .672 \end{pmatrix} \quad \mathbf{P}^4 = \begin{pmatrix} .328 & .672 \\ .332 & .668 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} .334 & .666 \\ .333 & .667 \end{pmatrix} \quad \mathbf{P}^6 = \begin{pmatrix} .333 & .667 \\ .333 & .667 \end{pmatrix} \quad \mathbf{P}^7 = \begin{pmatrix} .333 & .667 \\ .333 & .667 \end{pmatrix}$$

This sequence appears to be converging to a limiting matrix of the form

$$\mathbf{P}^\infty = \lim_{k\to\infty} \mathbf{P}^k = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix},$$

so the limiting population distribution is

$$\mathbf{p}_\infty^T = \lim_{k\to\infty} \mathbf{p}_k^T = \lim_{k\to\infty} \mathbf{p}_0^T \mathbf{T}^k = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{T}^k = \begin{pmatrix} n_0 & s_0 \end{pmatrix} \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{n_0 + s_0}{3} & \dfrac{2(n_0 + s_0)}{3} \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \end{pmatrix}.$$

Therefore, if the migration pattern continues to hold, then the population distribution will eventually stabilize with 1/3 of the population being in the North and 2/3 of the population in the South. And this is independent of the initial distribution! The powers of $\mathbf{P}$ indicate that the population distribution will be practically stable in no more than 6 years—individuals may continue to move, but the proportions in each region are essentially constant by the sixth year.

The operation of transposition has an interesting effect upon a matrix product—a reversal of order occurs.

Reverse Order Law for Transposition

For conformable matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T.$$

The case of conjugate transposition is similar. That is,

$$(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^*.$$

Proof. By definition,

$$(\mathbf{AB})_{ij}^T = [\mathbf{AB}]_{ji} = \mathbf{A}_{j*}\mathbf{B}_{*i}.$$

Consider the (i,j)-entry of the matrix $\mathbf{B}^T\mathbf{A}^T$ and write

$$\begin{aligned}\left[\mathbf{B}^T\mathbf{A}^T\right]_{ij} &= \left(\mathbf{B}^T\right)_{i*}\left(\mathbf{A}^T\right)_{*j} = \sum_k \left[\mathbf{B}^T\right]_{ik}\left[\mathbf{A}^T\right]_{kj} \\ &= \sum_k [\mathbf{B}]_{ki}[\mathbf{A}]_{jk} = \sum_k [\mathbf{A}]_{jk}[\mathbf{B}]_{ki} \\ &= \mathbf{A}_{j*}\mathbf{B}_{*i}.\end{aligned}$$

Therefore, $(\mathbf{AB})_{ij}^T = \left[\mathbf{B}^T\mathbf{A}^T\right]_{ij}$ for all i and j, and thus $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$. The proof for the conjugate transpose case is similar. ∎

Example 3.6.4

For every matrix $\mathbf{A}_{m\times n}$, the products $\mathbf{A}^T\mathbf{A}$ and $\mathbf{AA}^T$ are symmetric matrices because

$$\left(\mathbf{A}^T\mathbf{A}\right)^T = \mathbf{A}^T\mathbf{A}^{T^T} = \mathbf{A}^T\mathbf{A} \quad \text{and} \quad \left(\mathbf{AA}^T\right)^T = \mathbf{A}^{T^T}\mathbf{A}^T = \mathbf{AA}^T.$$

Example 3.6.5

Trace of a Product. Recall from Example 3.3.1 that the trace of a square matrix is the sum of its main diagonal entries. Although matrix multiplication is not commutative, the trace function is one of the few cases where the order of the matrices can be changed without affecting the results.

Problem: For matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times m}$, prove that

$$trace\,(\mathbf{AB}) = trace\,(\mathbf{BA}).$$

Solution:

$$\begin{aligned}trace\,(\mathbf{AB}) &= \sum_i [\mathbf{AB}]_{ii} = \sum_i \mathbf{A}_{i*}\mathbf{B}_{*i} = \sum_i\sum_k a_{ik}b_{ki} = \sum_i\sum_k b_{ki}a_{ik} \\ &= \sum_k\sum_i b_{ki}a_{ik} = \sum_k \mathbf{B}_{k*}\mathbf{A}_{*k} = \sum_k [\mathbf{BA}]_{kk} = trace\,(\mathbf{BA}).\end{aligned}$$

Note: This is true in spite of the fact that $\mathbf{AB}$ is $m\times m$ while $\mathbf{BA}$ is $n\times n$. Furthermore, this result can be extended to say that any product of conformable matrices can be permuted *cyclically* without altering the trace of the product. For example,

$$trace\,(\mathbf{ABC}) = trace\,(\mathbf{BCA}) = trace\,(\mathbf{CAB}).$$

However, a noncyclical permutation may not preserve the trace. For example,

$$trace\,(\mathbf{ABC}) \neq trace\,(\mathbf{BAC}).$$

Executing multiplication between two matrices by partitioning one or both factors into ***submatrices***—a matrix contained within another matrix—can be a useful technique.

Block Matrix Multiplication

Suppose that $\mathbf{A}$ and $\mathbf{B}$ are partitioned into submatrices—often referred to as ***blocks***—as indicated below.

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1r} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{s1} & \mathbf{A}_{s2} & \cdots & \mathbf{A}_{sr} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1t} \\ \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{r1} & \mathbf{B}_{r2} & \cdots & \mathbf{B}_{rt} \end{pmatrix}.$$

If the pairs $(\mathbf{A}_{ik}, \mathbf{B}_{kj})$ are conformable, then $\mathbf{A}$ and $\mathbf{B}$ are said to be ***conformably partitioned***. For such matrices, the product $\mathbf{AB}$ is formed by combining the blocks exactly the same way as the scalars are combined in ordinary matrix multiplication. That is, the (i,j)-block in $\mathbf{AB}$ is

$$\mathbf{A}_{i1}\mathbf{B}_{1j} + \mathbf{A}_{i2}\mathbf{B}_{2j} + \cdots + \mathbf{A}_{ir}\mathbf{B}_{rj}.$$

Although a completely general proof is possible, looking at some examples better serves the purpose of understanding this technique.

Example 3.6.6

Block multiplication is particularly useful when there are patterns in the matrices to be multiplied. Consider the partitioned matrices

$$\mathbf{A} = \left(\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right) = \begin{pmatrix} \mathbf{C} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1 & 2 & 1 & 2 \\ 3 & 4 & 3 & 4 \end{array}\right) = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{C} & \mathbf{C} \end{pmatrix},$$

where

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Using block multiplication, the product $\mathbf{AB}$ is easily computed to be

$$\mathbf{AB} = \begin{pmatrix} \mathbf{C} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{C} & \mathbf{C} \end{pmatrix} = \begin{pmatrix} 2\mathbf{C} & \mathbf{C} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} = \left(\begin{array}{cc|cc} 2 & 4 & 1 & 2 \\ 6 & 8 & 3 & 4 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right).$$

Example 3.6.7

Reducibility. Suppose that $\mathbf{T}_{n\times n}\mathbf{x} = \mathbf{b}$ represents a system of linear equations in which the coefficient matrix is ***block triangular***. That is, $\mathbf{T}$ can be partitioned as

$$\mathbf{T} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \quad \text{where} \quad \mathbf{A} \text{ is } r \times r \text{ and } \mathbf{C} \text{ is } n-r \times n-r. \tag{3.6.3}$$

If $\mathbf{x}$ and $\mathbf{b}$ are similarly partitioned as $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix}$, then block multiplication shows that $\mathbf{Tx} = \mathbf{b}$ reduces to two smaller systems

$$\begin{aligned} \mathbf{Ax}_1 + \mathbf{Bx}_2 &= \mathbf{b}_1, \\ \mathbf{Cx}_2 &= \mathbf{b}_2, \end{aligned}$$

so if all systems are consistent, a block version of back substitution is possible—i.e., solve $\mathbf{Cx}_2 = \mathbf{b}_2$ for $\mathbf{x}_2$, and substituted this back into $\mathbf{Ax}_1 = \mathbf{b}_1 - \mathbf{Bx}_2$, which is then solved for $\mathbf{x}_1$. For obvious reasons, block-triangular systems of this type are sometimes referred to as ***reducible systems,*** and $\mathbf{T}$ is said to be a ***reducible matrix.*** Recall that applying Gaussian elimination with back substitution to an $n \times n$ system requires about $n^3/3$ multiplications/divisions and about $n^3/3$ additions/subtractions. This means that it's more efficient to solve two smaller subsystems than to solve one large main system. For example, suppose the matrix $\mathbf{T}$ in (3.6.3) is 100×100 while $\mathbf{A}$ and $\mathbf{C}$ are each 50×50. If $\mathbf{Tx} = \mathbf{b}$ is solved without taking advantage of its reducibility, then about $10^6/3$ multiplications/divisions are needed. But by taking advantage of the reducibility, only about $(250 \times 10^3)/3$ multiplications/divisions are needed to solve both 50×50 subsystems. Another advantage of reducibility is realized when a computer's main memory capacity is not large enough to store the entire coefficient matrix but is large enough to hold the submatrices.

Exercises for section 3.6

3.6.1. For the partitioned matrices

$$\mathbf{A} = \left(\begin{array}{c|cc|ccc} 1 & 0 & 0 & 3 & 3 & 3 \\ 1 & 0 & 0 & 3 & 3 & 3 \\ \hline 1 & 2 & 2 & 0 & 0 & 0 \end{array}\right) \quad \text{and} \quad \mathbf{B} = \left(\begin{array}{cc} -1 & -1 \\ \hline 0 & 0 \\ 0 & 0 \\ \hline -1 & -2 \\ -1 & -2 \\ -1 & -2 \end{array}\right),$$

use block multiplication with the indicated partitions to form the product $\mathbf{AB}$.

3.6.2. For all matrices $\mathbf{A}_{n\times k}$ and $\mathbf{B}_{k\times n}$, show that the block matrix

$$\mathbf{L} = \begin{pmatrix} \mathbf{I} - \mathbf{BA} & \mathbf{B} \\ 2\mathbf{A} - \mathbf{ABA} & \mathbf{AB} - \mathbf{I} \end{pmatrix}$$

has the property $\mathbf{L}^2 = \mathbf{I}$. Matrices with this property are said to be ***involutory,*** and they occur in the science of cryptography.

3.6.3. For the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 1 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix},$$

determine $\mathbf{A}^{300}$. **Hint:** A square matrix $\mathbf{C}$ is said to be ***idempotent*** when it has the property that $\mathbf{C}^2 = \mathbf{C}$. Make use of idempotent submatrices in $\mathbf{A}$.

3.6.4. For every matrix $\mathbf{A}_{m\times n}$, demonstrate that the products $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$ are hermitian matrices.

3.6.5. If $\mathbf{A}$ and $\mathbf{B}$ are symmetric matrices that commute, prove that the product $\mathbf{AB}$ is also symmetric. If $\mathbf{AB} \neq \mathbf{BA}$, is $\mathbf{AB}$ necessarily symmetric?

3.6.6. Prove that the right-hand distributive property is true.

3.6.7. For each matrix $\mathbf{A}_{n\times n}$, explain why it is impossible to find a solution for $\mathbf{X}_{n\times n}$ in the matrix equation

$$\mathbf{AX} - \mathbf{XA} = \mathbf{I}.$$

Hint: Consider the trace function.

3.6.8. Let $\mathbf{y}^T_{1\times m}$ be a row of unknowns, and let $\mathbf{A}_{m\times n}$ and $\mathbf{b}^T_{1\times n}$ be known matrices.

(a) Explain why the matrix equation $\mathbf{y}^T\mathbf{A} = \mathbf{b}^T$ represents a system of n linear equations in m unknowns.

(b) How are the solutions for $\mathbf{y}^T$ in $\mathbf{y}^T\mathbf{A} = \mathbf{b}^T$ related to the solutions for $\mathbf{x}$ in $\mathbf{A}^T\mathbf{x} = \mathbf{b}$?

3.6.9. A particular electronic device consists of a collection of switching circuits that can be either in an ON state or an OFF state. These electronic switches are allowed to change state at regular time intervals called *clock cycles.* Suppose that at the end of each clock cycle, 30% of the switches currently in the OFF state change to ON, while 90% of those in the ON state revert to the OFF state.

(a) Show that the device approaches an equilibrium in the sense that the proportion of switches in each state eventually becomes constant, and determine these equilibrium proportions.

(b) Independent of the initial proportions, about how many clock cycles does it take for the device to become essentially stable?

3.6.10. Write the following system in the form $\mathbf{T}_{n\times n}\mathbf{x} = \mathbf{b}$, where $\mathbf{T}$ is block triangular, and then obtain the solution by solving two small systems as described in Example 3.6.7.

$$\begin{aligned} x_1 + x_2 + 3x_3 + 4x_4 &= -1, \\ 2x_3 + 3x_4 &= 3, \\ x_1 + 2x_2 + 5x_3 + 6x_4 &= -2, \\ x_3 + 2x_4 &= 4. \end{aligned}$$

3.6.11. Prove that each of the following statements is true for conformable matrices.

(a) $trace\,(\mathbf{ABC}) = trace\,(\mathbf{BCA}) = trace\,(\mathbf{CAB})$.

(b) $trace\,(\mathbf{ABC})$ can be different from $trace\,(\mathbf{BAC})$.

(c) $trace\left(\mathbf{A}^T\mathbf{B}\right) = trace\left(\mathbf{AB}^T\right)$.

3.6.12. Suppose that $\mathbf{A}_{m\times n}$ and $\mathbf{x}_{n\times 1}$ have real entries.

(a) Prove that $\mathbf{x}^T\mathbf{x} = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

(b) Prove that $trace\left(\mathbf{A}^T\mathbf{A}\right) = 0$ if and only if $\mathbf{A} = \mathbf{0}$.

3.7 MATRIX INVERSION

If α is a nonzero scalar, then for each number β the equation $\alpha x = \beta$ has a unique solution given by $x = \alpha^{-1}\beta$. To *prove* that $\alpha^{-1}\beta$ is a solution, write

$$\alpha(\alpha^{-1}\beta) = (\alpha\alpha^{-1})\beta = (1)\beta = \beta. \tag{3.7.1}$$

Uniqueness follows because if x_1 and x_2 are two solutions, then

$$\begin{aligned} \alpha x_1 = \beta = \alpha x_2 \quad &\Longrightarrow \quad \alpha^{-1}(\alpha x_1) = \alpha^{-1}(\alpha x_2) \\ &\Longrightarrow \quad (\alpha^{-1}\alpha)x_1 = (\alpha^{-1}\alpha)x_2 \\ &\Longrightarrow \quad (1)x_1 = (1)x_2 \quad \Longrightarrow \quad x_1 = x_2. \end{aligned} \tag{3.7.2}$$

These observations seem pedantic, but they are important in order to see how to make the transition from scalar equations to matrix equations. In particular, these arguments show that in addition to associativity, the properties

$$\alpha\alpha^{-1} = 1 \quad \text{and} \quad \alpha^{-1}\alpha = 1 \tag{3.7.3}$$

are the key ingredients, so if we want to solve matrix equations in the same fashion as we solve scalar equations, then a matrix analogue of (3.7.3) is needed.

Matrix Inversion

For a given square matrix $\mathbf{A}_{n\times n}$, the matrix $\mathbf{B}_{n\times n}$ that satisfies the conditions

$$\mathbf{AB} = \mathbf{I}_n \quad \text{and} \quad \mathbf{BA} = \mathbf{I}_n$$

is called the ***inverse*** of $\mathbf{A}$ and is denoted by $\mathbf{B} = \mathbf{A}^{-1}$. Not all square matrices are invertible—the zero matrix is a trivial example, but there are also many nonzero matrices that are not invertible. An invertible matrix is said to be ***nonsingular,*** and a square matrix with no inverse is called a ***singular matrix.***

Notice that matrix inversion is defined for square matrices only—the condition $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A}$ rules out inverses of nonsquare matrices.

Example 3.7.1

If

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \text{where} \quad \delta = ad - bc \neq 0,$$

then

$$\mathbf{A}^{-1} = \frac{1}{\delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

because it can be verified that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_2$.

Although not all matrices are invertible, *when an inverse exists, it is unique.* To see this, suppose that $\mathbf{X}_1$ and $\mathbf{X}_2$ are both inverses for a nonsingular matrix $\mathbf{A}$. Then

$$\mathbf{X}_1 = \mathbf{X}_1\mathbf{I} = \mathbf{X}_1(\mathbf{A}\mathbf{X}_2) = (\mathbf{X}_1\mathbf{A})\mathbf{X}_2 = \mathbf{I}\mathbf{X}_2 = \mathbf{X}_2,$$

which implies that only one inverse is possible.

Since matrix inversion was defined analogously to scalar inversion, and since matrix multiplication is associative, exactly the same reasoning used in (3.7.1) and (3.7.2) can be applied to a matrix equation $\mathbf{AX} = \mathbf{B}$, so we have the following statements.

Matrix Equations

- If $\mathbf{A}$ is a nonsingular matrix, then there is a unique solution for $\mathbf{X}$ in the matrix equation $\mathbf{A}_{n\times n}\mathbf{X}_{n\times p} = \mathbf{B}_{n\times p}$, and the solution is

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}. \tag{3.7.4}$$

- A system of n linear equations in n unknowns can be written as a single matrix equation $\mathbf{A}_{n\times n}\mathbf{x}_{n\times 1} = \mathbf{b}_{n\times 1}$ (see p. 99), so it follows from (3.7.4) that when $\mathbf{A}$ is nonsingular, the system has a unique solution given by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

However, it must be stressed that the representation of the solution as $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ is mostly a notational or theoretical convenience. *In practice, a nonsingular system* $\mathbf{Ax} = \mathbf{b}$ *is almost never solved by first computing* $\mathbf{A}^{-1}$ *and then the product* $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. The reason will be apparent when we learn how much work is involved in computing $\mathbf{A}^{-1}$.

Since not all square matrices are invertible, methods are needed to distinguish between nonsingular and singular matrices. There is a variety of ways to describe the class of nonsingular matrices, but those listed below are among the most important.

Existence of an Inverse

For an $n \times n$ matrix $\mathbf{A}$, the following statements are equivalent.

- $\mathbf{A}^{-1}$ exists ($\mathbf{A}$ is nonsingular). (3.7.5)
- $rank\,(\mathbf{A}) = n$. (3.7.6)
- $\mathbf{A} \xrightarrow{\text{Gauss–Jordan}} \mathbf{I}$. (3.7.7)
- $\mathbf{Ax} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$. (3.7.8)

Proof. The fact that (3.7.6) $\Longleftrightarrow$ (3.7.7) is a direct consequence of the definition of rank, and (3.7.6) $\Longleftrightarrow$ (3.7.8) was established in §2.4. Consequently, statements (3.7.6), (3.7.7), and (3.7.8) are equivalent, so if we establish that (3.7.5) $\Longleftrightarrow$ (3.7.6), then the proof will be complete.

Proof of (3.7.5) $\Longrightarrow$ (3.7.6). Begin by observing that (3.5.5) guarantees that a matrix $\mathbf{X} = [\mathbf{X}_{*1} \,|\, \mathbf{X}_{*2} \,|\, \cdots \,|\, \mathbf{X}_{*n}]$ satisfies the equation $\mathbf{AX} = \mathbf{I}$ if and only if $\mathbf{X}_{*j}$ is a solution of the linear system $\mathbf{Ax} = \mathbf{I}_{*j}$. If $\mathbf{A}$ is nonsingular, then we know from (3.7.4) that there exists a unique solution to $\mathbf{AX} = \mathbf{I}$, and hence each linear system $\mathbf{Ax} = \mathbf{I}_{*j}$ has a unique solution. But in §2.5 we learned that a linear system has a unique solution if and only if the rank of the coefficient matrix equals the number of unknowns, so $rank\,(\mathbf{A}) = n$.

Proof of (3.7.6) $\Longrightarrow$ (3.7.5). If $rank\,(\mathbf{A}) = n$, then (2.3.4) insures that each system $\mathbf{Ax} = \mathbf{I}_{*j}$ is consistent because $rank[\mathbf{A} \,|\, \mathbf{I}_{*j}] = n = rank\,(\mathbf{A})$. Furthermore, the results of §2.5 guarantee that each system $\mathbf{Ax} = \mathbf{I}_{*j}$ has a unique solution, and hence there is a unique solution to the matrix equation $\mathbf{AX} = \mathbf{I}$. We would like to say that $\mathbf{X} = \mathbf{A}^{-1}$, but we cannot jump to this conclusion without first arguing that $\mathbf{XA} = \mathbf{I}$. Suppose this is not true—i.e., suppose that $\mathbf{XA} - \mathbf{I} \neq \mathbf{0}$. Since

$$\mathbf{A}(\mathbf{XA} - \mathbf{I}) = (\mathbf{AX})\mathbf{A} - \mathbf{A} = \mathbf{IA} - \mathbf{A} = \mathbf{0},$$

it follows from (3.5.5) that any nonzero column of $\mathbf{XA} - \mathbf{I}$ is a nontrivial solution of the homogeneous system $\mathbf{Ax} = \mathbf{0}$. But this is a contradiction of the fact that (3.7.6) $\Longleftrightarrow$ (3.7.8). Therefore, the supposition that $\mathbf{XA} - \mathbf{I} \neq \mathbf{0}$ must be false, and thus $\mathbf{AX} = \mathbf{I} = \mathbf{XA}$, which means $\mathbf{A}$ is nonsingular. ∎

The definition of matrix inversion says that in order to compute $\mathbf{A}^{-1}$, it is necessary to solve *both* of the matrix equations $\mathbf{AX} = \mathbf{I}$ and $\mathbf{XA} = \mathbf{I}$. These two equations are necessary to rule out the possibility of nonsquare inverses. But when only square matrices are involved, then any one of the two equations will suffice—the following example elaborates.

Example 3.7.2

Problem: If $\mathbf{A}$ and $\mathbf{X}$ are *square* matrices, explain why

$$\mathbf{AX} = \mathbf{I} \quad \Longrightarrow \quad \mathbf{XA} = \mathbf{I}. \tag{3.7.9}$$

In other words, if $\mathbf{A}$ and $\mathbf{X}$ are square and $\mathbf{AX} = \mathbf{I}$, then $\mathbf{X} = \mathbf{A}^{-1}$.

Solution: Notice first that $\mathbf{AX} = \mathbf{I}$ implies $\mathbf{X}$ is nonsingular because if $\mathbf{X}$ is singular, then, by (3.7.8), there is a column vector $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{Xx} = \mathbf{0}$, which is contrary to the fact that $\mathbf{x} = \mathbf{Ix} = \mathbf{AXx} = \mathbf{0}$. Now that we know $\mathbf{X}^{-1}$ exists, we can establish (3.7.9) by writing

$$\mathbf{AX} = \mathbf{I} \quad \Longrightarrow \quad \mathbf{AXX}^{-1} = \mathbf{X}^{-1} \quad \Longrightarrow \quad \mathbf{A} = \mathbf{X}^{-1} \quad \Longrightarrow \quad \mathbf{XA} = \mathbf{I}.$$

Caution! The argument above is not valid for nonsquare matrices. When $m \neq n$, it's possible that $\mathbf{A}_{m\times n}\mathbf{X}_{n\times m} = \mathbf{I}_m$, but $\mathbf{XA} \neq \mathbf{I}_n$.

Although we usually try to avoid computing the inverse of a matrix, there are times when an inverse must be found. To construct an algorithm that will yield $\mathbf{A}^{-1}$ when $\mathbf{A}_{n\times n}$ is nonsingular, recall from Example 3.7.2 that determining $\mathbf{A}^{-1}$ is equivalent to solving the single matrix equation $\mathbf{AX} = \mathbf{I}$, and due to (3.5.5), this in turn is equivalent to solving the n linear systems defined by

$$\mathbf{Ax} = \mathbf{I}_{*j} \quad \text{for} \quad j = 1, 2, \ldots, n. \tag{3.7.10}$$

In other words, if $\mathbf{X}_{*1}, \mathbf{X}_{*2}, \ldots, \mathbf{X}_{*n}$ are the respective solutions to (3.7.10), then $\mathbf{X} = [\mathbf{X}_{*1} \,|\, \mathbf{X}_{*2} \,|\, \cdots \,|\, \mathbf{X}_{*n}]$ solves the equation $\mathbf{AX} = \mathbf{I}$, and hence $\mathbf{X} = \mathbf{A}^{-1}$. If $\mathbf{A}$ is nonsingular, then we know from (3.7.7) that the Gauss–Jordan method reduces the augmented matrix $[\mathbf{A} \,|\, \mathbf{I}_{*j}]$ to $[\mathbf{I} \,|\, \mathbf{X}_{*j}]$, and the results of §1.3 insure that $\mathbf{X}_{*j}$ is the unique solution to $\mathbf{Ax} = \mathbf{I}_{*j}$. That is,

$$[\mathbf{A} \,|\, \mathbf{I}_{*j}] \xrightarrow{\text{Gauss–Jordan}} \left[\mathbf{I} \,\middle|\, [\mathbf{A}^{-1}]_{*j}\right].$$

But rather than solving each system $\mathbf{Ax} = \mathbf{I}_{*j}$ separately, we can solve them simultaneously by taking advantage of the fact that they all have the same coefficient matrix. In other words, applying the Gauss–Jordan method to the larger augmented array $[\mathbf{A} \,|\, \mathbf{I}_{*1} \,|\, \mathbf{I}_{*2} \,|\, \cdots \,|\, \mathbf{I}_{*n}]$ produces

$$[\mathbf{A} \,|\, \mathbf{I}_{*1} \,|\, \mathbf{I}_{*2} \,|\, \cdots \,|\, \mathbf{I}_{*n}] \xrightarrow{\text{Gauss–Jordan}} \left[\mathbf{I} \,\middle|\, [\mathbf{A}^{-1}]_{*1} \,\middle|\, [\mathbf{A}^{-1}]_{*2} \,\middle|\, \cdots \,\middle|\, [\mathbf{A}^{-1}]_{*n}\right],$$

or more compactly,

$$[\mathbf{A} \,|\, \mathbf{I}] \xrightarrow{\text{Gauss–Jordan}} [\mathbf{I} \,|\, \mathbf{A}^{-1}]. \tag{3.7.11}$$

What happens if we try to invert a singular matrix using this procedure? The fact that (3.7.5) $\Longleftrightarrow$ (3.7.6) $\Longleftrightarrow$ (3.7.7) guarantees that a singular matrix $\mathbf{A}$ cannot be reduced to $\mathbf{I}$ by Gauss–Jordan elimination because a zero row will have to emerge in the left-hand side of the augmented array at some point during the process. This means that we do not need to know at the outset whether $\mathbf{A}$ is nonsingular or singular—it becomes self-evident depending on whether or not the reduction (3.7.11) can be completed. A summary is given below.

Computing an Inverse

Gauss–Jordan elimination can be used to invert $\mathbf{A}$ by the reduction

$$[\mathbf{A} \,|\, \mathbf{I}] \xrightarrow{\text{Gauss–Jordan}} [\mathbf{I} \,|\, \mathbf{A}^{-1}]. \tag{3.7.12}$$

The only way for this reduction to fail is for a row of zeros to emerge in the left-hand side of the augmented array, and this occurs if and only if $\mathbf{A}$ is a singular matrix. A different (and somewhat more practical) algorithm is given Example 3.10.3 on p. 148.

Although they are not included in the simple examples of this section, you are reminded that the pivoting and scaling strategies presented in §1.5 need to be incorporated, and the effects of ill-conditioning discussed in §1.6 must be considered whenever matrix inverses are computed using floating-point arithmetic. However, practical applications rarely require an inverse to be computed.

Example 3.7.3

Problem: If possible, find the inverse of $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$.

Solution:

$$[\mathbf{A}\,|\,\mathbf{I}] = \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 & 1 & 0 \\ 1 & 2 & 3 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|rrr} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|rrr} 1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|rrr} 1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 1 & 0 & -1 & 2 & -1 \\ 0 & 0 & 1 & 0 & -1 & 1 \end{array}\right)$$

Therefore, the matrix is nonsingular, and $\mathbf{A}^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$. If we wish to check this answer, we need only check that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. If this holds, then the result of Example 3.7.2 insures that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ will automatically be true.

Earlier in this section it was stated that one almost never solves a nonsingular linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ by first computing $\mathbf{A}^{-1}$ and then the product $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. To appreciate why this is true, pay attention to how much effort is required to perform one matrix inversion.

Operation Counts for Inversion

Computing $\mathbf{A}^{-1}_{n\times n}$ by reducing $[\mathbf{A}|\mathbf{I}]$ with Gauss–Jordan requires

- n^3 multiplications/divisions,
- $n^3 - 2n^2 + n$ additions/subtractions.

Interestingly, if Gaussian elimination with a back substitution process is applied to $[\mathbf{A}|\mathbf{I}]$ instead of the Gauss–Jordan technique, then exactly the same operation count can be obtained. Although Gaussian elimination with back substitution is more efficient than the Gauss–Jordan method for solving a single linear system, the two procedures are essentially equivalent for inversion.

Solving a nonsingular system $\mathbf{Ax} = \mathbf{b}$ by first computing $\mathbf{A}^{-1}$ and then forming the product $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ requires $n^3 + n^2$ multiplications/divisions and $n^3 - n^2$ additions/subtractions. Recall from §1.5 that Gaussian elimination with back substitution requires only about $n^3/3$ multiplications/divisions and about $n^3/3$ additions/subtractions. In other words, *using* $\mathbf{A}^{-1}$ *to solve a nonsingular system* $\mathbf{Ax} = \mathbf{b}$ *requires about three times the effort as does Gaussian elimination with back substitution.*

To put things in perspective, consider standard matrix multiplication between two $n \times n$ matrices. It is not difficult to verify that n^3 multiplications and $n^3 - n^2$ additions are required. Remarkably, it takes almost exactly as much effort to perform one matrix multiplication as to perform one matrix inversion. This fact always seems to be counter to a novice's intuition—it "feels" like matrix inversion should be a more difficult task than matrix multiplication, but this is not the case.

The remainder of this section is devoted to a discussion of some of the important properties of matrix inversion. We begin with the four basic facts listed below.

Properties of Matrix Inversion

For nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$, the following properties hold.

- $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$. (3.7.13)
- The product $\mathbf{AB}$ is also nonsingular. (3.7.14)
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ (the reverse order law for inversion). (3.7.15)
- $\left(\mathbf{A}^{-1}\right)^{T} = \left(\mathbf{A}^{T}\right)^{-1}$ and $\left(\mathbf{A}^{-1}\right)^{*} = (\mathbf{A}^{*})^{-1}$. (3.7.16)

Proof. Property (3.7.13) follows directly from the definition of inversion. To prove (3.7.14) and (3.7.15), let $\mathbf{X} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ and verify that $(\mathbf{AB})\mathbf{X} = \mathbf{I}$ by writing

$$(\mathbf{AB})\mathbf{X} = (\mathbf{AB})\mathbf{B}^{-1}\mathbf{A}^{-1} = \mathbf{A}(\mathbf{BB}^{-1})\mathbf{A}^{-1} = \mathbf{A}(\mathbf{I})\mathbf{A}^{-1} = \mathbf{AA}^{-1} = \mathbf{I}.$$

According to the discussion in Example 3.7.2, we are now guaranteed that $\mathbf{X}(\mathbf{AB}) = \mathbf{I}$, and we need not bother to verify it. To prove property (3.7.16), let $\mathbf{X} = \left(\mathbf{A}^{-1}\right)^{T}$ and verify that $\mathbf{A}^{T}\mathbf{X} = \mathbf{I}$. Make use of the reverse order law for transposition to write

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

Therefore, $\left(\mathbf{A}^{T}\right)^{-1} = \mathbf{X} = \left(\mathbf{A}^{-1}\right)^{T}$. The proof of the conjugate transpose case is similar. ∎

In general the product of two rank-r matrices does not necessarily have to produce another matrix of rank r. For example,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 2 & 4 \\ -1 & -2 \end{pmatrix}$$

each has rank 1, but the product $\mathbf{AB} = \mathbf{0}$ has rank 0. However, we saw in (3.7.14) that the product of two invertible matrices is again invertible. That is, if $rank\,(A_{n\times n}) = n$ and $rank\,(B_{n\times n}) = n$, then $rank\,(\mathbf{AB}) = n$. This generalizes to any number of matrices.

Products of Nonsingular Matrices Are Nonsingular

If $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ are each $n \times n$ nonsingular matrices, then the product $\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_k$ is also nonsingular, and its inverse is given by the reverse order law. That is,

$$(\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_k)^{-1} = \mathbf{A}_k^{-1}\cdots\mathbf{A}_2^{-1}\mathbf{A}_1^{-1}.$$

Proof. Apply (3.7.14) and (3.7.15) inductively. For example, when $k = 3$ you can write

$$(\mathbf{A}_1\{\mathbf{A}_2\mathbf{A}_3\})^{-1} = \{\mathbf{A}_2\mathbf{A}_3\}^{-1}\mathbf{A}_1^{-1} = \mathbf{A}_3^{-1}\mathbf{A}_2^{-1}\mathbf{A}_1^{-1}. \quad \blacksquare$$

Exercises for section 3.7

3.7.1. When possible, find the inverse of each of the following matrices. Check your answer by using matrix multiplication.

(a) $\begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ (c) $\begin{pmatrix} 4 & -8 & 5 \\ 4 & -7 & 4 \\ 3 & -4 & 2 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ (e) $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$

3.7.2. Find the matrix $\mathbf{X}$ such that $\mathbf{X} = \mathbf{AX} + \mathbf{B}$, where

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \end{pmatrix}.$$

3.7.3. For a square matrix $\mathbf{A}$, explain why each of the following statements must be true.

(a) If $\mathbf{A}$ contains a zero row or a zero column, then $\mathbf{A}$ is singular.

(b) If $\mathbf{A}$ contains two identical rows or two identical columns, then $\mathbf{A}$ is singular.

(c) If one row (or column) is a multiple of another row (or column), then $\mathbf{A}$ must be singular.

3.7.4. Answer each of the following questions.

(a) Under what conditions is a diagonal matrix nonsingular? Describe the structure of the inverse of a diagonal matrix.

(b) Under what conditions is a triangular matrix nonsingular? Describe the structure of the inverse of a triangular matrix.

3.7.5. If $\mathbf{A}$ is nonsingular and symmetric, prove that $\mathbf{A}^{-1}$ is symmetric.

3.7.6. If $\mathbf{A}$ is a square matrix such that $\mathbf{I}-\mathbf{A}$ is nonsingular, prove that

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

3.7.7. Prove that if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times m$ such that $\mathbf{AB} = \mathbf{I}_m$ and $\mathbf{BA} = \mathbf{I}_n$, then $m = n$.

3.7.8. If $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{A}+\mathbf{B}$ are each nonsingular, prove that

$$\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = \left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}.$$

3.7.9. Let $\mathbf{S}$ be a skew-symmetric matrix with real entries.

(a) Prove that $\mathbf{I}-\mathbf{S}$ is nonsingular. **Hint:** $\mathbf{x}^T\mathbf{x} = 0 \implies \mathbf{x} = \mathbf{0}$.

(b) If $\mathbf{A} = (\mathbf{I}+\mathbf{S})(\mathbf{I}-\mathbf{S})^{-1}$, show that $\mathbf{A}^{-1} = \mathbf{A}^T$.

3.7.10. For matrices $\mathbf{A}_{r\times r}$, $\mathbf{B}_{s\times s}$, and $\mathbf{C}_{r\times s}$ such that $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, verify that each of the following is true.

(a) $$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{-1} \end{pmatrix}$$

(b) $$\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{B}^{-1} \\ \mathbf{0} & \mathbf{B}^{-1} \end{pmatrix}$$

3.7.11. Consider the block matrix $\begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{C}_{r\times s} \\ \mathbf{R}_{s\times r} & \mathbf{B}_{s\times s} \end{pmatrix}$. When the indicated inverses exist, the matrices defined by

$$\mathbf{S} = \mathbf{B} - \mathbf{R}\mathbf{A}^{-1}\mathbf{C} \quad \text{and} \quad \mathbf{T} = \mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{R}$$

are called the ***Schur complements***[20] of $\mathbf{A}$ and $\mathbf{B}$, respectively.

(a) If $\mathbf{A}$ and $\mathbf{S}$ are both nonsingular, verify that

$$\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{R} & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{C}\mathbf{S}^{-1}\mathbf{R}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{S}^{-1} \\ -\mathbf{S}^{-1}\mathbf{R}\mathbf{A}^{-1} & \mathbf{S}^{-1} \end{pmatrix}.$$

(b) If $\mathbf{B}$ and $\mathbf{T}$ are nonsingular, verify that

$$\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{R} & \mathbf{B} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{T}^{-1} & -\mathbf{T}^{-1}\mathbf{C}\mathbf{B}^{-1} \\ -\mathbf{B}^{-1}\mathbf{R}\mathbf{T}^{-1} & \mathbf{B}^{-1} + \mathbf{B}^{-1}\mathbf{R}\mathbf{T}^{-1}\mathbf{C}\mathbf{B}^{-1} \end{pmatrix}.$$

3.7.12. Suppose that $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are $n \times n$ matrices such that $\mathbf{A}\mathbf{B}^T$ and $\mathbf{C}\mathbf{D}^T$ are each symmetric and $\mathbf{A}\mathbf{D}^T - \mathbf{B}\mathbf{C}^T = \mathbf{I}$. Prove that

$$\mathbf{A}^T\mathbf{D} - \mathbf{C}^T\mathbf{B} = \mathbf{I}.$$

[20] This is named in honor of the German mathematician Issai Schur (1875–1941), who first studied matrices of this type. Schur was a student and collaborator of Ferdinand Georg Frobenius (p. 662). Schur and Frobenius were among the first to study matrix theory as a discipline unto itself, and each made great contributions to the subject. It was Emilie V. Haynsworth (1916–1987)—a mathematical granddaughter of Schur—who introduced the phrase "Schur complement" and developed several important aspects of the concept.

3.8 INVERSES OF SUMS AND SENSITIVITY

The reverse order law for inversion makes the inverse of a product easy to deal with, but the inverse of a sum is much more difficult. To begin with, $(\mathbf{A}+\mathbf{B})^{-1}$ may not exist even if $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ each exist. Moreover, if $(\mathbf{A}+\mathbf{B})^{-1}$ exists, then, with rare exceptions, $(\mathbf{A}+\mathbf{B})^{-1} \neq \mathbf{A}^{-1}+\mathbf{B}^{-1}$. This doesn't even hold for scalars (i.e., 1×1 matrices), so it has no chance of holding in general.

There is no useful general formula for $(\mathbf{A}+\mathbf{B})^{-1}$, but there are some special sums for which something can be said. One of the most easily inverted sums is $\mathbf{I}+\mathbf{c}\mathbf{d}^T$ in which $\mathbf{c}$ and $\mathbf{d}$ are $n\times 1$ nonzero columns such that $1+\mathbf{d}^T\mathbf{c}\neq 0$. It's straightforward to verify by direct multiplication that

$$\left(\mathbf{I}+\mathbf{c}\mathbf{d}^T\right)^{-1} = \mathbf{I} - \frac{\mathbf{c}\mathbf{d}^T}{1+\mathbf{d}^T\mathbf{c}}. \tag{3.8.1}$$

If $\mathbf{I}$ is replaced by a nonsingular matrix $\mathbf{A}$ satisfying $1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}\neq 0$, then the reverse order law for inversion in conjunction with (3.8.1) yields

$$\begin{aligned}(\mathbf{A}+\mathbf{c}\mathbf{d}^T)^{-1} &= \Big(\mathbf{A}(\mathbf{I}+\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T)\Big)^{-1} = (\mathbf{I}+\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T)^{-1}\mathbf{A}^{-1}\\ &= \left(\mathbf{I} - \frac{\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T}{1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}}\right)\mathbf{A}^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\mathbf{A}^{-1}}{1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}}.\end{aligned}$$

This is often called the Sherman–Morrison[21] rank-one update formula because it can be shown (Exercise 3.9.9, p. 140) that $rank\,(\mathbf{c}\mathbf{d}^T)=1$ when $\mathbf{c}\neq\mathbf{0}\neq\mathbf{d}$.

Sherman–Morrison Formula

- If $\mathbf{A}_{n\times n}$ is nonsingular and if $\mathbf{c}$ and $\mathbf{d}$ are $n\times 1$ columns such that $1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}\neq 0$, then the sum $\mathbf{A}+\mathbf{c}\mathbf{d}^T$ is nonsingular, and

$$\left(\mathbf{A}+\mathbf{c}\mathbf{d}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\mathbf{A}^{-1}}{1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}}. \tag{3.8.2}$$

- The *Sherman–Morrison–Woodbury formula* is a generalization. If $\mathbf{C}$ and $\mathbf{D}$ are $n\times k$ such that $(\mathbf{I}+\mathbf{D}^T\mathbf{A}^{-1}\mathbf{C})^{-1}$ exists, then

$$(\mathbf{A}+\mathbf{C}\mathbf{D}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}(\mathbf{I}+\mathbf{D}^T\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{D}^T\mathbf{A}^{-1}. \tag{3.8.3}$$

[21] This result appeared in the 1949–1950 work of American statisticians J. Sherman and W. J. Morrison, but they were not the first to discover it. The formula was independently presented by the English mathematician W. J. Duncan in 1944 and by American statisticians L. Guttman (1946), Max Woodbury (1950), and M. S. Bartlett (1951). Since its derivation is so natural, it almost certainly was discovered by many others along the way. Recognition and fame are often not afforded simply for introducing an idea, but rather for applying the idea to a useful end.

The Sherman–Morrison–Woodbury formula (3.8.3) can be verified with direct multiplication, or it can be derived as indicated in Exercise 3.8.6.

To appreciate the utility of the Sherman–Morrison formula, suppose $\mathbf{A}^{-1}$ is known from a previous calculation, but now one entry in $\mathbf{A}$ needs to be changed or updated—say we need to add α to a_{ij}. It's not necessary to start from scratch to compute the new inverse because Sherman–Morrison shows how the previously computed information in $\mathbf{A}^{-1}$ can be updated to produce the new inverse. Let $\mathbf{c} = \mathbf{e}_i$ and $\mathbf{d} = \alpha\mathbf{e}_j$, where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the ith and j^{th} unit columns, respectively. The matrix $\mathbf{c}\mathbf{d}^T$ has α in the (i,j)-position and zeros elsewhere so that

$$\mathbf{B} = \mathbf{A} + \mathbf{c}\mathbf{d}^T = \mathbf{A} + \alpha\mathbf{e}_i\mathbf{e}_j^T$$

is the updated matrix. According to the Sherman–Morrison formula,

$$\begin{aligned}\mathbf{B}^{-1} &= \left(\mathbf{A} + \alpha\mathbf{e}_i\mathbf{e}_j^T\right)^{-1} = \mathbf{A}^{-1} - \alpha\frac{\mathbf{A}^{-1}\mathbf{e}_i\mathbf{e}_j^T\mathbf{A}^{-1}}{1+\alpha\mathbf{e}_j^T\mathbf{A}^{-1}\mathbf{e}_i} \\ &= \mathbf{A}^{-1} - \alpha\frac{[\mathbf{A}^{-1}]_{*i}[\mathbf{A}^{-1}]_{j*}}{1+\alpha[\mathbf{A}^{-1}]_{ji}} \quad \text{(recall Exercise 3.5.4).}\end{aligned} \tag{3.8.4}$$

This shows how $\mathbf{A}^{-1}$ changes when a_{ij} is perturbed, and it provides a useful algorithm for updating $\mathbf{A}^{-1}$.

Example 3.8.1

Problem: Start with $\mathbf{A}$ and $\mathbf{A}^{-1}$ given below. Update $\mathbf{A}$ by adding 1 to a_{21}, and then use the Sherman–Morrison formula to update $\mathbf{A}^{-1}$:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^{-1} = \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix}.$$

Solution: The updated matrix is

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \end{pmatrix} = \mathbf{A} + \mathbf{e}_2\mathbf{e}_1^T.$$

Applying the Sherman–Morrison formula yields the updated inverse

$$\begin{aligned}\mathbf{B}^{-1} &= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{e}_2\mathbf{e}_1^T\mathbf{A}^{-1}}{1+\mathbf{e}_1^T\mathbf{A}^{-1}\mathbf{e}_2} = \mathbf{A}^{-1} - \frac{[\mathbf{A}^{-1}]_{*2}[\mathbf{A}^{-1}]_{1*}}{1+[\mathbf{A}^{-1}]_{12}} \\ &= \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix} - \frac{\begin{pmatrix} -2 \\ 1 \end{pmatrix}\begin{pmatrix} 3 & -2 \end{pmatrix}}{1-2} = \begin{pmatrix} -3 & 2 \\ 2 & -1 \end{pmatrix}.\end{aligned}$$

Another sum that often requires inversion is $\mathbf{I}-\mathbf{A}$, but we have to be careful because $(\mathbf{I}-\mathbf{A})^{-1}$ need not always exist. However, we are safe when the entries in $\mathbf{A}$ are sufficiently small. In particular, if the entries in $\mathbf{A}$ are small enough in magnitude to insure that $\lim_{n\to\infty}\mathbf{A}^n=\mathbf{0}$, then, analogous to scalar algebra,

$$(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{n-1})=\mathbf{I}-\mathbf{A}^n\to\mathbf{I}\quad\text{as}\quad n\to\infty,$$

so we have the following matrix version of a geometric series.

Neumann Series

If $\lim_{n\to\infty}\mathbf{A}^n=\mathbf{0}$, then $\mathbf{I}-\mathbf{A}$ is nonsingular and

$$(\mathbf{I}-\mathbf{A})^{-1}=\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots=\sum_{k=0}^{\infty}\mathbf{A}^k. \tag{3.8.5}$$

This is the ***Neumann series.*** It provides approximations of $(\mathbf{I}-\mathbf{A})^{-1}$ when $\mathbf{A}$ has entries of small magnitude. For example, a first-order approximation is $(\mathbf{I}-\mathbf{A})^{-1}\approx\mathbf{I}+\mathbf{A}$. More on the Neumann series appears in Example 7.3.1, p. 527, and the complete statement is developed on p. 618.

While there is no useful formula for $(\mathbf{A}+\mathbf{B})^{-1}$ in general, the Neumann series allows us to say something when $\mathbf{B}$ has small entries relative to $\mathbf{A}$, or vice versa. For example, if $\mathbf{A}^{-1}$ exists, and if the entries in $\mathbf{B}$ are small enough in magnitude to insure that $\lim_{n\to\infty}\left(\mathbf{A}^{-1}\mathbf{B}\right)^n=\mathbf{0}$, then

$$\begin{aligned}(\mathbf{A}+\mathbf{B})^{-1}&=\Big(\mathbf{A}\left(\mathbf{I}-\left[-\mathbf{A}^{-1}\mathbf{B}\right]\right)\Big)^{-1}=\Big(\mathbf{I}-\left[-\mathbf{A}^{-1}\mathbf{B}\right]\Big)^{-1}\mathbf{A}^{-1}\\&=\left(\sum_{k=0}^{\infty}\left[-\mathbf{A}^{-1}\mathbf{B}\right]^k\right)\mathbf{A}^{-1},\end{aligned}$$

and a first-order approximation is

$$(\mathbf{A}+\mathbf{B})^{-1}\approx\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}. \tag{3.8.6}$$

Consequently, if $\mathbf{A}$ is perturbed by a small matrix $\mathbf{B}$, possibly resulting from errors due to inexact measurements or perhaps from roundoff error, then the resulting change in $\mathbf{A}^{-1}$ is about $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. In other words, the effect of a small perturbation (or error) $\mathbf{B}$ is magnified by multiplication (on both sides) with $\mathbf{A}^{-1}$, so if $\mathbf{A}^{-1}$ has large entries, small perturbations (or errors) in $\mathbf{A}$ can produce large perturbations (or errors) in the resulting inverse. You can reach

essentially the same conclusion from (3.8.4) when only a single entry is perturbed and from Exercise 3.8.2 when a single column is perturbed.

This discussion resolves, at least in part, an issue raised in §1.6—namely, "What mechanism determines the extent to which a nonsingular system $\mathbf{Ax} = \mathbf{b}$ is ill-conditioned?" To see how, an aggregate measure of the magnitude of the entries in $\mathbf{A}$ is needed, and one common measure is

$$\|\mathbf{A}\| = \max_i \sum_j |a_{ij}| = \text{the maximum absolute row sum.} \tag{3.8.7}$$

This is one example of a *matrix norm,* a detailed discussion of which is given in §5.1. Theoretical properties specific to (3.8.7) are developed on pp. 280 and 283, and one property established there is the fact that $\|\mathbf{XY}\| \leq \|\mathbf{X}\|\,\|\mathbf{Y}\|$ for all conformable matrices $\mathbf{X}$ and $\mathbf{Y}$. But let's keep things on an intuitive level for the time being and defer the details. Using the norm (3.8.7), the approximation (3.8.6) insures that if $\|\mathbf{B}\|$ is sufficiently small, then

$$\left\|\mathbf{A}^{-1} - (\mathbf{A}+\mathbf{B})^{-1}\right\| \approx \left\|\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right\| \leq \left\|\mathbf{A}^{-1}\right\|\,\|\mathbf{B}\|\,\left\|\mathbf{A}^{-1}\right\|,$$

so, if we interpret $x \lesssim y$ to mean that x is bounded above by something not far from y, we can write

$$\frac{\left\|\mathbf{A}^{-1} - (\mathbf{A}+\mathbf{B})^{-1}\right\|}{\|\mathbf{A}^{-1}\|} \lesssim \left\|\mathbf{A}^{-1}\right\|\,\|\mathbf{B}\| = \left\|\mathbf{A}^{-1}\right\|\,\|\mathbf{A}\|\left\{\frac{\|\mathbf{B}\|}{\|\mathbf{A}\|}\right\}.$$

The term on the left is the relative change in the inverse, and $\|\mathbf{B}\| \,/\, \|\mathbf{A}\|$ is the relative change in $\mathbf{A}$. The number $\kappa = \left\|\mathbf{A}^{-1}\right\|\,\|\mathbf{A}\|$ is therefore the "magnification factor" that dictates how much the relative change in $\mathbf{A}$ is magnified. This magnification factor κ is called a ***condition number*** for $\mathbf{A}$. In other words, if κ is small relative to 1 (i.e., if $\mathbf{A}$ is *well conditioned*), then a small relative change (or error) in $\mathbf{A}$ cannot produce a large relative change (or error) in the inverse, but if κ is large (i.e., if $\mathbf{A}$ is *ill conditioned*), then a small relative change (or error) in $\mathbf{A}$ can possibly (but not necessarily) result in a large relative change (or error) in the inverse.

The situation for linear systems is similar. If the coefficients in a nonsingular system $\mathbf{Ax} = \mathbf{b}$ are slightly perturbed to produce the system $(\mathbf{A}+\mathbf{B})\tilde{\mathbf{x}} = \mathbf{b}$, then $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ and $\tilde{\mathbf{x}} = (\mathbf{A}+\mathbf{B})^{-1}\mathbf{b}$ so that (3.8.6) implies

$$\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{b} - (\mathbf{A}+\mathbf{B})^{-1}\mathbf{b} \approx \mathbf{A}^{-1}\mathbf{b} - \left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right)\mathbf{b} = \mathbf{A}^{-1}\mathbf{B}\mathbf{x}.$$

For column vectors, (3.8.7) reduces to $\|\mathbf{x}\| = \max_i |x_i|$, and we have

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \lesssim \left\|\mathbf{A}^{-1}\right\|\,\|\mathbf{B}\|\,\|\mathbf{x}\|,$$

so the relative change in the solution is

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \lesssim \left\|\mathbf{A}^{-1}\right\| \|\mathbf{B}\| = \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| \left\{\frac{\|\mathbf{B}\|}{\|\mathbf{A}\|}\right\} = \kappa \left\{\frac{\|\mathbf{B}\|}{\|\mathbf{A}\|}\right\}. \tag{3.8.8}$$

Again, the condition number κ is pivotal because when κ is small, a small relative change in $\mathbf{A}$ cannot produce a large relative change in $\mathbf{x}$, but for larger values of κ, a small relative change in $\mathbf{A}$ can possibly result in a large relative change in $\mathbf{x}$. Below is a summary of these observations.

Sensitivity and Conditioning

- A nonsingular matrix $\mathbf{A}$ is said to be ***ill conditioned*** if a small relative change in $\mathbf{A}$ can cause a large relative change in $\mathbf{A}^{-1}$. The degree of ill-conditioning is gauged by a ***condition number*** $\kappa = \|\mathbf{A}\| \|\mathbf{A}^{-1}\|$, where $\|\star\|$ is a matrix norm.
- The sensitivity of the solution of $\mathbf{Ax} = \mathbf{b}$ to perturbations (or errors) in $\mathbf{A}$ is measured by the extent to which $\mathbf{A}$ is an ill-conditioned matrix. More is said in Example 5.12.1 on p. 414.

Example 3.8.2

It was demonstrated in Example 1.6.1 that the system

$$\begin{aligned} .835x + .667y &= .168, \\ .333x + .266y &= .067, \end{aligned}$$

is sensitive to small perturbations. We can understand this in the current context by examining the condition number of the coefficient matrix. If the matrix norm (3.8.7) is employed with

$$\mathbf{A} = \begin{pmatrix} .835 & .667 \\ .333 & .266 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^{-1} = \begin{pmatrix} -266000 & 667000 \\ 333000 & -835000 \end{pmatrix},$$

then the condition number for $\mathbf{A}$ is

$$\kappa = \kappa = \|\mathbf{A}\| \|\mathbf{A}^{-1}\| = (1.502)(1168000) = 1,754,336 \approx 1.7 \times 10^6.$$

Since the right-hand side of (3.8.8) is only an estimate of the relative error in the solution, the exact value of κ is not as important as its order of magnitude. Because κ is of order 10^6, (3.8.8) holds the possibility that the relative change (or error) in the solution can be about a million times larger than the relative

change (or error) in $\mathbf{A}$. Therefore, we must consider $\mathbf{A}$ and the associated linear system to be ill conditioned.

A Rule of Thumb. If Gaussian elimination with partial pivoting is used to solve a well-scaled nonsingular system $\mathbf{Ax} = \mathbf{b}$ using t-digit floating-point arithmetic, then, assuming no other source of error exists, it can be argued that when κ is of order 10^p, the computed solution is expected to be accurate to at least $t - p$ significant digits, more or less. In other words, one expects to lose roughly p significant figures. For example, if Gaussian elimination with 8-digit arithmetic is used to solve the 2×2 system given above, then only about $t - p = 8 - 6 = 2$ significant figures of accuracy should be expected. This doesn't preclude the possibility of getting lucky and attaining a higher degree of accuracy—it just says that you shouldn't bet the farm on it.

The complete story of conditioning has not yet been told. As pointed out earlier, it's about three times more costly to compute $\mathbf{A}^{-1}$ than to solve $\mathbf{Ax} = \mathbf{b}$, so it doesn't make sense to compute $\mathbf{A}^{-1}$ just to estimate the condition of $\mathbf{A}$. Questions concerning condition estimation without explicitly computing an inverse still need to be addressed. Furthermore, liberties allowed by using the $\approx$ and $\lesssim$ symbols produce results that are intuitively correct but not rigorous. Rigor will eventually be attained—see Example 5.12.1on p. 414.

Exercises for section 3.8

3.8.1. Suppose you are given that

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & -1 \\ -1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 2 \end{pmatrix}.$$

(a) Use the Sherman–Morrison formula to determine the inverse of the matrix $\mathbf{B}$ that is obtained by changing the $(3,2)$-entry in $\mathbf{A}$ from 0 to 2.

(b) Let $\mathbf{C}$ be the matrix that agrees with $\mathbf{A}$ except that $c_{32} = 2$ and $c_{33} = 2$. Use the Sherman–Morrison formula to find $\mathbf{C}^{-1}$.

3.8.2. Suppose $\mathbf{A}$ and $\mathbf{B}$ are nonsingular matrices in which $\mathbf{B}$ is obtained from $\mathbf{A}$ by replacing $\mathbf{A}_{*j}$ with another column $\mathbf{b}$. Use the Sherman–Morrison formula to derive the fact that

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} - \frac{\left(\mathbf{A}^{-1}\mathbf{b} - \mathbf{e}_j\right)[\mathbf{A}^{-1}]_{j*}}{[\mathbf{A}^{-1}]_{j*}\mathbf{b}}.$$

3.8.3. Suppose the coefficient matrix of a nonsingular system $\mathbf{Ax} = \mathbf{b}$ is updated to produce another nonsingular system $(\mathbf{A} + \mathbf{cd}^T)\mathbf{z} = \mathbf{b}$, where $\mathbf{b}, \mathbf{c}, \mathbf{d} \in \Re^{n\times 1}$, and let $\mathbf{y}$ be the solution of $\mathbf{Ay} = \mathbf{c}$. Show that $\mathbf{z} = \mathbf{x} - \mathbf{yd}^T\mathbf{x}/(1 + \mathbf{d}^T\mathbf{y})$.

3.8.4. (a) Use the Sherman–Morrison formula to prove that if $\mathbf{A}$ is nonsingular, then $\mathbf{A} + \alpha\mathbf{e}_i\mathbf{e}_j^T$ is nonsingular for a sufficiently small α.

(b) Use part (a) to prove that $\mathbf{I} + \mathbf{E}$ is nonsingular when all ϵ_{ij} 's are sufficiently small in magnitude. This is an alternative to using the Neumann series argument.

3.8.5. For given matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A}$ is nonsingular, explain why $\mathbf{A} + \epsilon\mathbf{B}$ is also nonsingular when the real number ϵ is constrained to a sufficiently small interval about the origin. In other words, prove that small perturbations of nonsingular matrices are also nonsingular.

3.8.6. Derive the Sherman–Morrison–Woodbury formula. **Hint:** Recall Exercise 3.7.11, and consider the product $\begin{pmatrix} \mathbf{I} & \mathbf{C} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{D}^T & -\mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{D}^T & \mathbf{I} \end{pmatrix}$.

3.8.7. Using the norm (3.8.7), rank the following matrices according to their degree of ill-conditioning:

$$\mathbf{A} = \begin{pmatrix} 100 & 0 & -100 \\ 0 & 100 & -100 \\ -100 & -100 & 300 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 8 & -1 \\ -9 & -71 & 11 \\ 1 & 17 & 18 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} 1 & 22 & -42 \\ 0 & 1 & -45 \\ -45 & -948 & 1 \end{pmatrix}.$$

3.8.8. Suppose that the entries in $\mathbf{A}(t)$, $\mathbf{x}(t)$, and $\mathbf{b}(t)$ are differentiable functions of a real variable t such that $\mathbf{A}(t)\mathbf{x}(t) = \mathbf{b}(t)$.

(a) Assuming that $\mathbf{A}(t)^{-1}$ exists, explain why

$$\frac{d\mathbf{A}(t)^{-1}}{dt} = -\mathbf{A}(t)^{-1}\mathbf{A}'(t)\mathbf{A}(t)^{-1}.$$

(b) Derive the equation

$$\mathbf{x}'(t) = \mathbf{A}(t)^{-1}\mathbf{b}'(t) - \mathbf{A}(t)^{-1}\mathbf{A}'(t)\mathbf{x}(t).$$

This shows that $\mathbf{A}^{-1}$ magnifies both the change in $\mathbf{A}$ and the change in $\mathbf{b}$, and thus it confirms the observation derived from (3.8.8) saying that the sensitivity of a nonsingular system to small perturbations is directly related to the magnitude of the entries in $\mathbf{A}^{-1}$.

3.9 ELEMENTARY MATRICES AND EQUIVALENCE

A common theme in mathematics is to break complicated objects into more elementary components, such as factoring large polynomials into products of smaller polynomials. The purpose of this section is to lay the groundwork for similar ideas in matrix algebra by considering how a general matrix might be factored into a product of more "elementary" matrices.

Elementary Matrices

Matrices of the form $\mathbf{I}-\mathbf{u}\mathbf{v}^T$, where $\mathbf{u}$ and $\mathbf{v}$ are $n \times 1$ columns such that $\mathbf{v}^T\mathbf{u} \neq 1$ are called ***elementary matrices***, and we know from (3.8.1) that all such matrices are nonsingular and

$$\left(\mathbf{I}-\mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{I} - \frac{\mathbf{u}\mathbf{v}^T}{\mathbf{v}^T\mathbf{u}-1}. \tag{3.9.1}$$

Notice that inverses of elementary matrices are elementary matrices.

We are primarily interested in the elementary matrices associated with the three elementary row (or column) operations hereafter referred to as follows.

- Type I is interchanging rows (columns) i and j.
- Type II is multiplying row (column) i by $\alpha \neq 0$.
- Type III is adding a multiple of row (column) i to row (column) j.

An elementary matrix of Type I, II, or III is created by performing an elementary operation of Type I, II, or III to an identity matrix. For example, the matrices

$$\mathbf{E}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{E}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{E}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \alpha & 0 & 1 \end{pmatrix} \tag{3.9.2}$$

are elementary matrices of Types I, II, and III, respectively, because $\mathbf{E}_1$ arises by interchanging rows 1 and 2 in $\mathbf{I}_3$, whereas $\mathbf{E}_2$ is generated by multiplying row 2 in $\mathbf{I}_3$ by α, and $\mathbf{E}_3$ is constructed by multiplying row 1 in $\mathbf{I}_3$ by α and adding the result to row 3. The matrices in (3.9.2) also can be generated by column operations. For example, $\mathbf{E}_3$ can be obtained by adding α times the third column of $\mathbf{I}_3$ to the first column. The fact that $\mathbf{E}_1$, $\mathbf{E}_2$, and $\mathbf{E}_3$ are of the form (3.9.1) follows by using the unit columns $\mathbf{e}_i$ to write

$$\mathbf{E}_1 = \mathbf{I}-\mathbf{u}\mathbf{u}^T, \text{ where } \mathbf{u} = \mathbf{e}_1-\mathbf{e}_2, \quad \mathbf{E}_2 = \mathbf{I}-(1-\alpha)\mathbf{e}_2\mathbf{e}_2^T, \quad \text{and} \quad \mathbf{E}_3 = \mathbf{I}+\alpha\mathbf{e}_3\mathbf{e}_1^T.$$

These observations generalize to matrices of arbitrary size.

One of our objectives is to remove the arrows from Gaussian elimination because the inability to do "arrow algebra" limits the theoretical analysis. For example, while it makes sense to add two equations together, there is no meaningful analog for arrows—reducing $\mathbf{A} \to \mathbf{B}$ and $\mathbf{C} \to \mathbf{D}$ by row operations does not guarantee that $\mathbf{A} + \mathbf{C} \to \mathbf{B} + \mathbf{D}$ is possible. The following properties are the mechanisms needed to remove the arrows from elimination processes.

Properties of Elementary Matrices

- When used as a *left-hand* multiplier, an elementary matrix of Type I, II, or III executes the corresponding *row* operation.
- When used as a *right-hand* multiplier, an elementary matrix of Type I, II, or III executes the corresponding *column* operation.

Proof. A proof for Type III operations is given—the other two cases are left to the reader. Using $\mathbf{I} + \alpha \mathbf{e}_j \mathbf{e}_i^T$ as a left-hand multiplier on an arbitrary matrix $\mathbf{A}$ produces

$$\left(\mathbf{I} + \alpha \mathbf{e}_j \mathbf{e}_i^T\right) \mathbf{A} = \mathbf{A} + \alpha \mathbf{e}_j \mathbf{A}_{i*} = \mathbf{A} + \alpha \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \leftarrow j^{th} \text{ row}.$$

This is exactly the matrix produced by a Type III row operation in which the ith row of $\mathbf{A}$ is multiplied by α and added to the jth row. When $\mathbf{I} + \alpha \mathbf{e}_j \mathbf{e}_i^T$ is used as a right-hand multiplier on $\mathbf{A}$, the result is

$$\mathbf{A}\left(\mathbf{I} + \alpha \mathbf{e}_j \mathbf{e}_i^T\right) = \mathbf{A} + \alpha \mathbf{A}_{*j} \mathbf{e}_i^T = \mathbf{A} + \alpha \begin{array}{c} \begin{array}{ccccc} & & i^{th} \text{ col} & & \\ & & \downarrow & & \end{array} \\ \begin{pmatrix} 0 & \cdots & a_{1j} & \cdots & 0 \\ 0 & \cdots & a_{2j} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & a_{nj} & \cdots & 0 \end{pmatrix} \end{array}.$$

This is the result of a Type III column operation in which the jth column of $\mathbf{A}$ is multiplied by α and then added to the ith column. ∎

Example 3.9.1

The sequence of row operations used to reduce $\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 4 & 8 \\ 3 & 6 & 13 \end{pmatrix}$ to $\mathbf{E_A}$ is indicated below.

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 4 & 8 \\ 3 & 6 & 13 \end{pmatrix} \begin{matrix} R_2 - 2R_1 \\ R_3 - 3R_1 \end{matrix} \longrightarrow \begin{pmatrix} 1 & 2 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\xrightarrow{\textit{Interchange } R_2 \textit{ and } R_3} \begin{pmatrix} 1 & 2 & 4 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} R_1 - 4R_2 \longrightarrow \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{E_A}.$$

The reduction can be accomplished by a sequence of left-hand multiplications with the corresponding elementary matrices as shown below.

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

The product of these elementary matrices is $\mathbf{P} = \begin{pmatrix} 13 & 0 & -4 \\ -3 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}$, and you can verify that it is indeed the case that $\mathbf{PA} = \mathbf{E_A}$. Thus the arrows are eliminated by replacing them with a product of elementary matrices.

We are now in a position to understand why nonsingular matrices are precisely those matrices that can be factored as a product of elementary matrices.

Products of Elementary Matrices

- $\mathbf{A}$ is a nonsingular matrix if and only if $\mathbf{A}$ is the product of elementary matrices of Type I, II, or III. (3.9.3)

Proof. If $\mathbf{A}$ is nonsingular, then the Gauss–Jordan technique reduces $\mathbf{A}$ to $\mathbf{I}$ by row operations. If $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k$ is the sequence of elementary matrices that corresponds to the elementary row operations used, then

$$\mathbf{G}_k \cdots \mathbf{G}_2 \mathbf{G}_1 \mathbf{A} = \mathbf{I} \text{ or, equivalently, } \mathbf{A} = \mathbf{G}_1^{-1} \mathbf{G}_2^{-1} \cdots \mathbf{G}_k^{-1}.$$

Since the inverse of an elementary matrix is again an elementary matrix of the same type, this proves that $\mathbf{A}$ is the product of elementary matrices of Type I, II, or III. Conversely, if $\mathbf{A} = \mathbf{E}_1 \mathbf{E}_2 \cdots \mathbf{E}_k$ is a product of elementary matrices, then $\mathbf{A}$ must be nonsingular because the $\mathbf{E}_i$'s are nonsingular, and a product of nonsingular matrices is also nonsingular. ∎

Equivalence

- Whenever $\mathbf{B}$ can be derived from $\mathbf{A}$ by a combination of elementary row and column operations, we write $\mathbf{A} \sim \mathbf{B}$, and we say that $\mathbf{A}$ and $\mathbf{B}$ are ***equivalent matrices***. Since elementary row and column operations are left-hand and right-hand multiplication by elementary matrices, respectively, and in view of (3.9.3), we can say that
$$\mathbf{A} \sim \mathbf{B} \iff \mathbf{PAQ} = \mathbf{B} \quad \text{for nonsingular } \mathbf{P} \text{ and } \mathbf{Q}.$$
- Whenever $\mathbf{B}$ can be obtained from $\mathbf{A}$ by performing a sequence of elementary *row* operations only, we write $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$, and we say that $\mathbf{A}$ and $\mathbf{B}$ are ***row equivalent***. In other words,
$$\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \iff \mathbf{PA} = \mathbf{B} \quad \text{for a nonsingular } \mathbf{P}.$$
- Whenever $\mathbf{B}$ can be obtained from $\mathbf{A}$ by performing a sequence of *column* operations only, we write $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$, and we say that $\mathbf{A}$ and $\mathbf{B}$ are ***column equivalent***. In other words,
$$\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B} \iff \mathbf{AQ} = \mathbf{B} \quad \text{for a nonsingular } \mathbf{Q}.$$

If it's possible to go from $\mathbf{A}$ to $\mathbf{B}$ by elementary row and column operations, then clearly it's possible to start with $\mathbf{B}$ and get back to $\mathbf{A}$ because elementary operations are reversible—i.e., $\mathbf{PAQ} = \mathbf{B} \implies \mathbf{P}^{-1}\mathbf{BQ}^{-1} = \mathbf{A}$. It therefore makes sense to talk about the equivalence of a pair of matrices without regard to order. In other words, $\mathbf{A} \sim \mathbf{B} \iff \mathbf{B} \sim \mathbf{A}$. Furthermore, it's not difficult to see that each type of equivalence is ***transitive*** in the sense that
$$\mathbf{A} \sim \mathbf{B} \quad \text{and} \quad \mathbf{B} \sim \mathbf{C} \implies \mathbf{A} \sim \mathbf{C}.$$

In §2.2 it was stated that each matrix $\mathbf{A}$ possesses a unique reduced row echelon form $\mathbf{E_A}$, and we accepted this fact because it is intuitively evident. However, we are now in a position to understand a rigorous proof.

Example 3.9.2

Problem: Prove that $\mathbf{E_A}$ is uniquely determined by $\mathbf{A}$.

Solution: Without loss of generality, we may assume that $\mathbf{A}$ is square—otherwise the appropriate number of zero rows or columns can be adjoined to $\mathbf{A}$ without affecting the results. Suppose that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E}_1$ and $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E}_2$, where $\mathbf{E}_1$ and $\mathbf{E}_2$ are both in reduced row echelon form. Consequently, $\mathbf{E}_1 \overset{\text{row}}{\sim} \mathbf{E}_2$, and hence there is a nonsingular matrix $\mathbf{P}$ such that
$$\mathbf{PE}_1 = \mathbf{E}_2. \tag{3.9.4}$$

Furthermore, by permuting the rows of $\mathbf{E}_1$ and $\mathbf{E}_2$ to force the pivotal 1's to occupy the diagonal positions, we see that

$$\mathbf{E}_1 \overset{\text{row}}{\sim} \mathbf{T}_1 \quad \text{and} \quad \mathbf{E}_2 \overset{\text{row}}{\sim} \mathbf{T}_2, \tag{3.9.5}$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are upper-triangular matrices in which the basic columns in each $\mathbf{T}_i$ occupy the same positions as the basic columns in $\mathbf{E}_i$. For example, if

$$\mathbf{E} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \text{ then } \mathbf{T} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Each $\mathbf{T}_i$ has the property that $\mathbf{T}_i^2 = \mathbf{T}_i$ because there is a ***permutation matrix*** $\mathbf{Q}_i$ (a product of elementary interchange matrices of Type I) such that

$$\mathbf{Q}_i\mathbf{T}_i\mathbf{Q}_i^T = \begin{pmatrix} \mathbf{I}_{r_i} & \mathbf{J}_i \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \text{ or, equivalently, } \mathbf{T}_i = \mathbf{Q}_i^T \begin{pmatrix} \mathbf{I}_{r_i} & \mathbf{J}_i \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}_i,$$

and $\mathbf{Q}_i^T = \mathbf{Q}_i^{-1}$ (see Exercise 3.9.4) implies $\mathbf{T}_i^2 = \mathbf{T}_i$. It follows from (3.9.5) that $\mathbf{T}_1 \overset{\text{row}}{\sim} \mathbf{T}_2$, so there is a nonsingular matrix $\mathbf{R}$ such that $\mathbf{R}\mathbf{T}_1 = \mathbf{T}_2$. Thus

$$\mathbf{T}_2 = \mathbf{R}\mathbf{T}_1 = \mathbf{R}\mathbf{T}_1\mathbf{T}_1 = \mathbf{T}_2\mathbf{T}_1 \quad \text{and} \quad \mathbf{T}_1 = \mathbf{R}^{-1}\mathbf{T}_2 = \mathbf{R}^{-1}\mathbf{T}_2\mathbf{T}_2 = \mathbf{T}_1\mathbf{T}_2.$$

Because $\mathbf{T}_1$ and $\mathbf{T}_2$ are both upper triangular, $\mathbf{T}_1\mathbf{T}_2$ and $\mathbf{T}_2\mathbf{T}_1$ have the same diagonal entries, and hence $\mathbf{T}_1$ and $\mathbf{T}_2$ have the same diagonal. Therefore, the positions of the basic columns (i.e., the pivotal positions) in $\mathbf{T}_1$ agree with those in $\mathbf{T}_2$, and hence $\mathbf{E}_1$ and $\mathbf{E}_2$ have basic columns in exactly the same positions. This means there is a permutation matrix $\mathbf{Q}$ such that

$$\mathbf{E}_1\mathbf{Q} = \begin{pmatrix} \mathbf{I}_r & \mathbf{J}_1 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{E}_2\mathbf{Q} = \begin{pmatrix} \mathbf{I}_r & \mathbf{J}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Using (3.9.4) yields $\mathbf{P}\mathbf{E}_1\mathbf{Q} = \mathbf{E}_2\mathbf{Q}$, or

$$\begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_r & \mathbf{J}_1 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_r & \mathbf{J}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

which in turn implies that $\mathbf{P}_{11} = \mathbf{I}_r$ and $\mathbf{P}_{11}\mathbf{J}_1 = \mathbf{J}_2$. Consequently, $\mathbf{J}_1 = \mathbf{J}_2$, and it follows that $\mathbf{E}_1 = \mathbf{E}_2$.

In passing, notice that the uniqueness of $\mathbf{E}_{\mathbf{A}}$ implies the uniqueness of the pivot positions in any other row echelon form derived from $\mathbf{A}$. If $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{U}_1$ and $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{U}_2$, where $\mathbf{U}_1$ and $\mathbf{U}_2$ are row echelon forms with different pivot positions, then Gauss–Jordan reduction applied to $\mathbf{U}_1$ and $\mathbf{U}_2$ would lead to two different reduced echelon forms, which is impossible.

In §2.2 we observed the fact that the column relationships in a matrix $\mathbf{A}$ are exactly the same as the column relationships in $\mathbf{E}_{\mathbf{A}}$. This observation is a special case of the more general result presented below.

Column and Row Relationships

- If $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$, then linear relationships existing among columns of $\mathbf{A}$ also hold among corresponding columns of $\mathbf{B}$. That is,

$$\mathbf{B}_{*k} = \sum_{j=1}^{n} \alpha_j \mathbf{B}_{*j} \quad \text{if and only if} \quad \mathbf{A}_{*k} = \sum_{j=1}^{n} \alpha_j \mathbf{A}_{*j}. \tag{3.9.6}$$

- In particular, the column relationships in $\mathbf{A}$ and $\mathbf{E_A}$ must be identical, so the nonbasic columns in $\mathbf{A}$ must be linear combinations of the basic columns in $\mathbf{A}$ as described in (2.2.3).
- If $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$, then linear relationships existing among rows of $\mathbf{A}$ must also hold among corresponding rows of $\mathbf{B}$.
- **Summary.** *Row* equivalence preserves *column* relationships, and *column* equivalence preserves *row* relationships.

Proof. If $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$, then $\mathbf{PA} = \mathbf{B}$ for some nonsingular $\mathbf{P}$. Recall from (3.5.5) that the jth column in $\mathbf{B}$ is given by

$$\mathbf{B}_{*j} = (\mathbf{PA})_{*j} = \mathbf{PA}_{*j}.$$

Therefore, if $\mathbf{A}_{*k} = \sum_j \alpha_j \mathbf{A}_{*j}$, then multiplication by $\mathbf{P}$ on the left produces $\mathbf{B}_{*k} = \sum_j \alpha_j \mathbf{B}_{*j}$. Conversely, if $\mathbf{B}_{*k} = \sum_j \alpha_j \mathbf{B}_{*j}$, then multiplication on the left by $\mathbf{P}^{-1}$ produces $\mathbf{A}_{*k} = \sum_j \alpha_j \mathbf{A}_{*j}$. The statement concerning column equivalence follows by considering transposes. ∎

The reduced row echelon form $\mathbf{E_A}$ is as far as we can go in reducing $\mathbf{A}$ by using only row operations. However, if we are allowed to use row operations in conjunction with column operations, then, as described below, the end result of a complete reduction is much simpler.

Rank Normal Form

If $\mathbf{A}$ is an $m \times n$ matrix such that $rank\,(\mathbf{A}) = r$, then

$$\mathbf{A} \sim \mathbf{N}_r = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{3.9.7}$$

$\mathbf{N}_r$ is called the ***rank normal form*** for $\mathbf{A}$, and it is the end product of a complete reduction of $\mathbf{A}$ by using both row and column operations.

Proof. It is always true that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E_A}$ so that there is a nonsingular matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{E_A}$. If $rank\,(\mathbf{A}) = r$, then the basic columns in $\mathbf{E_A}$ are the r unit columns. Apply column interchanges to $\mathbf{E_A}$ so as to move these r unit columns to the far left-hand side. If $\mathbf{Q}_1$ is the product of the elementary matrices corresponding to these column interchanges, then $\mathbf{PAQ}_1$ has the form

$$\mathbf{PAQ}_1 = \mathbf{E_A Q}_1 = \begin{pmatrix} \mathbf{I}_r & \mathbf{J} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Multiplying both sides of this equation on the right by the nonsingular matrix

$$\mathbf{Q}_2 = \begin{pmatrix} \mathbf{I}_r & -\mathbf{J} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \text{ produces } \mathbf{PAQ}_1\mathbf{Q}_2 = \begin{pmatrix} \mathbf{I}_r & \mathbf{J} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{I}_r & -\mathbf{J} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Thus $\mathbf{A} \sim \mathbf{N}_r$. because $\mathbf{P}$ and $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$ are nonsingular. ∎

Example 3.9.3

Problem: Explain why $rank\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} = rank\,(\mathbf{A}) + rank\,(\mathbf{B})$.

Solution: If $rank\,(\mathbf{A}) = r$ and $rank\,(\mathbf{B}) = s$, then $\mathbf{A} \sim \mathbf{N}_r$ and $\mathbf{B} \sim \mathbf{N}_s$. Consequently,

$$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \sim \begin{pmatrix} \mathbf{N}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_s \end{pmatrix} \implies rank\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} = rank\begin{pmatrix} \mathbf{N}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_s \end{pmatrix} = r + s.$$

Given matrices $\mathbf{A}$ and $\mathbf{B}$, how do we decide whether or not $\mathbf{A} \sim \mathbf{B}$, $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$, or $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$? We could use a trial-and-error approach by attempting to reduce $\mathbf{A}$ to $\mathbf{B}$ by elementary operations, but this would be silly because there are easy tests, as described below.

Testing for Equivalence

For $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ the following statements are true.

- $\mathbf{A} \sim \mathbf{B}$ if and only if $rank\,(\mathbf{A}) = rank\,(\mathbf{B})$. (3.9.8)
- $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$ if and only if $\mathbf{E_A} = \mathbf{E_B}$. (3.9.9)
- $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$ if and only if $\mathbf{E}_{\mathbf{A}^T} = \mathbf{E}_{\mathbf{B}^T}$. (3.9.10)

Corollary. Multiplication by nonsingular matrices cannot change rank.

Proof. To establish the validity of (3.9.8), observe that $rank\,(\mathbf{A}) = rank\,(\mathbf{B})$ implies $\mathbf{A} \sim \mathbf{N}_r$ and $\mathbf{B} \sim \mathbf{N}_r$. Therefore, $\mathbf{A} \sim \mathbf{N}_r \sim \mathbf{B}$. Conversely, if $\mathbf{A} \sim \mathbf{B}$, where $rank\,(\mathbf{A}) = r$ and $rank\,(\mathbf{B}) = s$, then $\mathbf{A} \sim \mathbf{N}_r$ and $\mathbf{B} \sim \mathbf{N}_s$, and hence $\mathbf{N}_r \sim \mathbf{A} \sim \mathbf{B} \sim \mathbf{N}_s$. Clearly, $\mathbf{N}_r \sim \mathbf{N}_s$ implies $r = s$. To prove (3.9.9), suppose first that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$. Because $\mathbf{B} \overset{\text{row}}{\sim} \mathbf{E_B}$, it follows that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E_B}$. Since a matrix has a uniquely determined reduced echelon form, it must be the case that $\mathbf{E_B} = \mathbf{E_A}$. Conversely, if $\mathbf{E_A} = \mathbf{E_B}$, then

$$\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E_A} = \mathbf{E_B} \overset{\text{row}}{\sim} \mathbf{B} \implies \mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}.$$

The proof of (3.9.10) follows from (3.9.9) by considering transposes because

$$\begin{aligned} \mathbf{A} \overset{\text{col}}{\sim} \mathbf{B} &\iff \mathbf{AQ} = \mathbf{B} \iff (\mathbf{AQ})^T = \mathbf{B}^T \\ &\iff \mathbf{Q}^T\mathbf{A}^T = \mathbf{B}^T \iff \mathbf{A}^T \overset{\text{row}}{\sim} \mathbf{B}^T. \quad \blacksquare \end{aligned}$$

Example 3.9.4

Problem: Are the relationships that exist among the columns in $\mathbf{A}$ the same as the column relationships in $\mathbf{B}$, and are the row relationships in $\mathbf{A}$ the same as the row relationships in $\mathbf{B}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ -4 & -3 & -1 \\ 2 & 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ 2 & 1 & -1 \end{pmatrix}?$$

Solution: Straightforward computation reveals that

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

and hence $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$. Therefore, the column relationships in $\mathbf{A}$ and $\mathbf{B}$ must be identical, and they must be the same as those in $\mathbf{E_A}$. Examining $\mathbf{E_A}$ reveals that $\mathbf{E}_{*3} = -2\mathbf{E}_{*1} + 3\mathbf{E}_{*2}$, so it must be the case that

$$\mathbf{A}_{*3} = -2\mathbf{A}_{*1} + 3\mathbf{A}_{*2} \quad \text{and} \quad \mathbf{B}_{*3} = -2\mathbf{B}_{*1} + 3\mathbf{B}_{*2}.$$

The row relationships in $\mathbf{A}$ and $\mathbf{B}$ are different because $\mathbf{E}_{\mathbf{A}^T} \neq \mathbf{E}_{\mathbf{B}^T}$.

On the surface, it may not seem plausible that a matrix and its transpose should have the same rank. After all, if $\mathbf{A}$ is 3×100, then $\mathbf{A}$ can have as many as 100 basic columns, but $\mathbf{A}^T$ can have at most three. Nevertheless, we can now demonstrate that $rank\,(\mathbf{A}) = rank\,(\mathbf{A}^T)$.

Transposition and Rank

Transposition does not change the rank—i.e., for all $m \times n$ matrices,

$$rank\,(\mathbf{A}) = rank\left(\mathbf{A}^T\right) \quad \text{and} \quad rank\,(\mathbf{A}) = rank\,(\mathbf{A}^*). \tag{3.9.11}$$

Proof. Let $rank\,(\mathbf{A}) = r$, and let $\mathbf{P}$ and $\mathbf{Q}$ be nonsingular matrices such that

$$\mathbf{PAQ} = \mathbf{N}_r = \begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times n-r} \\ \mathbf{0}_{m-r\times r} & \mathbf{0}_{m-r\times n-r} \end{pmatrix}.$$

Applying the reverse order law for transposition produces $\mathbf{Q}^T\mathbf{A}^T\mathbf{P}^T = \mathbf{N}_r^T$. Since $\mathbf{Q}^T$ and $\mathbf{P}^T$ are nonsingular, it follows that $\mathbf{A}^T \sim \mathbf{N}_r^T$, and therefore

$$rank\left(\mathbf{A}^T\right) = rank\left(\mathbf{N}_r^T\right) = rank\begin{pmatrix} \mathbf{I}_r & \mathbf{0}_{r\times m-r} \\ \mathbf{0}_{n-r\times r} & \mathbf{0}_{n-r\times m-r} \end{pmatrix} = r = rank\,(\mathbf{A}).$$

To prove $rank\,(\mathbf{A}) = rank\,(\mathbf{A}^*)$, write $\mathbf{N}_r = \overline{\mathbf{N}_r} = \overline{\mathbf{PAQ}} = \bar{\mathbf{P}}\bar{\mathbf{A}}\bar{\mathbf{Q}}$, and use the fact that the conjugate of a nonsingular matrix is again nonsingular (because $\bar{\mathbf{K}}^{-1} = \overline{\mathbf{K}^{-1}}$) to conclude that $\mathbf{N}_r \sim \overline{\mathbf{A}}$, and hence $rank\,(\mathbf{A}) = rank\left(\bar{\mathbf{A}}\right)$. It now follows from $rank\,(\mathbf{A}) = rank\left(\mathbf{A}^T\right)$ that

$$rank\,(\mathbf{A}^*) = rank\left(\bar{\mathbf{A}}^T\right) = rank\left(\bar{\mathbf{A}}\right) = rank\,(\mathbf{A}). \quad \blacksquare$$

Exercises for section 3.9

3.9.1. Suppose that $\mathbf{A}$ is an $m \times n$ matrix.

(a) If $[\mathbf{A}|\mathbf{I}_m]$ is row reduced to a matrix $[\mathbf{B}|\mathbf{P}]$, explain why $\mathbf{P}$ must be a nonsingular matrix such that $\mathbf{PA} = \mathbf{B}$.

(b) If $\left[\frac{\mathbf{A}}{\mathbf{I}_n}\right]$ is column reduced to $\left[\frac{\mathbf{C}}{\mathbf{Q}}\right]$, explain why $\mathbf{Q}$ must be a nonsingular matrix such that $\mathbf{AQ} = \mathbf{C}$.

(c) Find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{E}_{\mathbf{A}}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 7 \\ 1 & 2 & 3 & 6 \end{pmatrix}.$$

(d) Find nonsingular matrices $\mathbf{P}$ and $\mathbf{Q}$ such that $\mathbf{PAQ}$ is in rank normal form.

3.9.2. Consider the two matrices

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 0 & -1 \\ 3 & -1 & 4 & 0 \\ 0 & -8 & 8 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 2 & -6 & 8 & 2 \\ 5 & 1 & 4 & -1 \\ 3 & -9 & 12 & 3 \end{pmatrix}.$$

(a) Are $\mathbf{A}$ and $\mathbf{B}$ equivalent?
(b) Are $\mathbf{A}$ and $\mathbf{B}$ row equivalent?
(c) Are $\mathbf{A}$ and $\mathbf{B}$ column equivalent?

3.9.3. If $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$, explain why the basic columns in $\mathbf{A}$ occupy exactly the same positions as the basic columns in $\mathbf{B}$.

3.9.4. A product of elementary interchange matrices—i.e., elementary matrices of Type I—is called a ***permutation matrix.*** If $\mathbf{P}$ is a permutation matrix, explain why $\mathbf{P}^{-1} = \mathbf{P}^T$.

3.9.5. If $\mathbf{A}_{n\times n}$ is a nonsingular matrix, which (if any) of the following statements are true?

(a) $\mathbf{A} \sim \mathbf{A}^{-1}$. (b) $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{A}^{-1}$. (c) $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{A}^{-1}$.
(d) $\mathbf{A} \sim \mathbf{I}$. (e) $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{I}$. (f) $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{I}$.

3.9.6. Which (if any) of the following statements are true?

(a) $\mathbf{A} \sim \mathbf{B} \implies \mathbf{A}^T \sim \mathbf{B}^T$. (b) $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \implies \mathbf{A}^T \overset{\text{row}}{\sim} \mathbf{B}^T$.
(c) $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \implies \mathbf{A}^T \overset{\text{col}}{\sim} \mathbf{B}^T$. (d) $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \implies \mathbf{A} \sim \mathbf{B}$.
(e) $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B} \implies \mathbf{A} \sim \mathbf{B}$. (f) $\mathbf{A} \sim \mathbf{B} \implies \mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$.

3.9.7. Show that every elementary matrix of Type I can be written as a product of elementary matrices of Types II and III. **Hint:** Recall Exercise 1.2.12 on p. 14.

3.9.8. If $rank\,(\mathbf{A}_{m\times n}) = r$, show that there exist matrices $\mathbf{B}_{m\times r}$ and $\mathbf{C}_{r\times n}$ such that $\mathbf{A} = \mathbf{BC}$, where $rank\,(\mathbf{B}) = rank\,(\mathbf{C}) = r$. Such a factorization is called a ***full-rank factorization.*** **Hint:** Consider the basic columns of $\mathbf{A}$ and the nonzero rows of $\mathbf{E_A}$.

3.9.9. Prove that $rank\,(\mathbf{A}_{m\times n}) = 1$ if and only if there are nonzero columns $\mathbf{u}_{m\times 1}$ and $\mathbf{v}_{n\times 1}$ such that

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T.$$

3.9.10. Prove that if $rank\,(\mathbf{A}_{n\times n}) = 1$, then $\mathbf{A}^2 = \tau\mathbf{A}$, where $\tau = trace\,(\mathbf{A})$.

3.10 THE LU FACTORIZATION

We have now come full circle, and we are back to where the text began—solving a nonsingular system of linear equations using Gaussian elimination with back substitution. This time, however, the goal is to describe and understand the process in the context of matrices.

If $\mathbf{Ax} = \mathbf{b}$ is a nonsingular system, then the object of Gaussian elimination is to reduce $\mathbf{A}$ to an upper-triangular matrix using elementary row operations. If no zero pivots are encountered, then row interchanges are not necessary, and the reduction can be accomplished by using only elementary row operations of Type III. For example, consider reducing the matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 7 & 7 \\ 6 & 18 & 22 \end{pmatrix}$$

to upper-triangular form as shown below:

$$\begin{aligned} \begin{pmatrix} 2 & 2 & 2 \\ 4 & 7 & 7 \\ 6 & 18 & 22 \end{pmatrix} \begin{matrix} \\ R_2 - 2R_1 \\ R_3 - 3R_1 \end{matrix} &\longrightarrow \begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 12 & 16 \end{pmatrix} \begin{matrix} \\ \\ R_3 - 4R_2 \end{matrix} \\ &\longrightarrow \begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{pmatrix} = \mathbf{U}. \end{aligned} \tag{3.10.1}$$

We learned in the previous section that each of these Type III operations can be executed by means of a left-hand multiplication with the corresponding elementary matrix $\mathbf{G}_i$, and the product of all of these $\mathbf{G}_i$'s is

$$\mathbf{G}_3\mathbf{G}_2\mathbf{G}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 5 & -4 & 1 \end{pmatrix}.$$

In other words, $\mathbf{G}_3\mathbf{G}_2\mathbf{G}_1\mathbf{A} = \mathbf{U}$, so that $\mathbf{A} = \mathbf{G}_1^{-1}\mathbf{G}_2^{-1}\mathbf{G}_3^{-1}\mathbf{U} = \mathbf{LU}$, where $\mathbf{L}$ is the lower-triangular matrix

$$\mathbf{L} = \mathbf{G}_1^{-1}\mathbf{G}_2^{-1}\mathbf{G}_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix}.$$

Thus $\mathbf{A} = \mathbf{LU}$ is a product of a lower-triangular matrix $\mathbf{L}$ and an upper-triangular matrix $\mathbf{U}$. Naturally, this is called an $\mathbf{LU}$ ***factorization*** of $\mathbf{A}$.

Observe that $\mathbf{U}$ is the end product of Gaussian elimination and has the pivots on its diagonal, while $\mathbf{L}$ has 1's on its diagonal. Moreover, $\mathbf{L}$ has the remarkable property that below its diagonal, *each entry* ℓ_{ij} *is precisely the multiplier used in the elimination* (3.10.1) *to annihilate the* (i,j)*-position.*

This is characteristic of what happens in general. To develop the general theory, it's convenient to introduce the concept of an ***elementary lower-triangular matrix,*** which is defined to be an $n \times n$ triangular matrix of the form

$$\mathbf{T}_k = \mathbf{I} - \mathbf{c}_k \mathbf{e}_k^T,$$

where $\mathbf{c}_k$ is a column with zeros in the first k positions. In particular, if

$$\mathbf{c}_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \mu_{k+1} \\ \vdots \\ \mu_n \end{pmatrix}, \quad \text{then} \quad \mathbf{T}_k = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -\mu_{k+1} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\mu_n & 0 & \cdots & 1 \end{pmatrix}. \tag{3.10.2}$$

By observing that $\mathbf{e}_k^T \mathbf{c}_k = 0$, the formula for the inverse of an elementary matrix given in (3.9.1) produces

$$\mathbf{T}_k^{-1} = \mathbf{I} + \mathbf{c}_k \mathbf{e}_k^T = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \mu_{k+1} & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_n & 0 & \cdots & 1 \end{pmatrix}, \tag{3.10.3}$$

which is also an elementary lower-triangular matrix. The utility of elementary lower-triangular matrices lies in the fact that all of the Type III row operations needed to annihilate the entries below the kth pivot can be accomplished with one multiplication by $\mathbf{T}_k$. If

$$\mathbf{A}_{k-1} = \begin{pmatrix} * & * & \cdots & \alpha_1 & * & \cdots & * \\ 0 & * & \cdots & \alpha_2 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \alpha_k & * & \cdots & * \\ 0 & 0 & \cdots & \alpha_{k+1} & * & \cdots & * \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_n & * & \cdots & * \end{pmatrix}$$

is the partially triangularized result after $k-1$ elimination steps, then

$$\mathbf{T}_k\mathbf{A}_{k-1} = \left(\mathbf{I} - \mathbf{c}_k\mathbf{e}_k^T\right)\mathbf{A}_{k-1} = \mathbf{A}_{k-1} - \mathbf{c}_k\mathbf{e}_k^T\mathbf{A}_{k-1}$$

$$= \begin{pmatrix} * & * & \cdots & \alpha_1 & * & \cdots & * \\ 0 & * & \cdots & \alpha_2 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \alpha_k & * & \cdots & * \\ 0 & 0 & \cdots & 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & * & \cdots & * \end{pmatrix}, \qquad \text{where} \quad \mathbf{c}_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \alpha_{k+1}/\alpha_k \\ \vdots \\ \alpha_n/\alpha_k \end{pmatrix}$$

contains the multipliers used to annihilate those entries below α_k. Notice that $\mathbf{T}_k$ does not alter the first $k-1$ columns of $\mathbf{A}_{k-1}$ because $\mathbf{e}_k^T\left[\mathbf{A}_{k-1}\right]_{*j} = 0$ whenever $j \le k-1$. Therefore, if no row interchanges are required, then reducing $\mathbf{A}$ to an upper-triangular matrix $\mathbf{U}$ by Gaussian elimination is equivalent to executing a sequence of $n-1$ left-hand multiplications with elementary lower-triangular matrices. That is, $\mathbf{T}_{n-1}\cdots\mathbf{T}_2\mathbf{T}_1\mathbf{A} = \mathbf{U}$, and hence

$$\mathbf{A} = \mathbf{T}_1^{-1}\mathbf{T}_2^{-1}\cdots\mathbf{T}_{n-1}^{-1}\mathbf{U}. \tag{3.10.4}$$

Making use of the fact that $\mathbf{e}_j^T\mathbf{c}_k = 0$ whenever $j \le k$ and applying (3.10.3) reveals that

$$\begin{aligned} \mathbf{T}_1^{-1}\mathbf{T}_2^{-1}\cdots\mathbf{T}_{n-1}^{-1} &= \left(\mathbf{I} + \mathbf{c}_1\mathbf{e}_1^T\right)\left(\mathbf{I} + \mathbf{c}_2\mathbf{e}_2^T\right)\cdots\left(\mathbf{I} + \mathbf{c}_{n-1}\mathbf{e}_{n-1}^T\right) \\ &= \mathbf{I} + \mathbf{c}_1\mathbf{e}_1^T + \mathbf{c}_2\mathbf{e}_2^T + \cdots + \mathbf{c}_{n-1}\mathbf{e}_{n-1}^T. \end{aligned} \tag{3.10.5}$$

By observing that

$$\mathbf{c}_k\mathbf{e}_k^T = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \ell_{k+1,k} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \ell_{nk} & 0 & \cdots & 0 \end{pmatrix},$$

where the ℓ_{ik}'s are the multipliers used at the kth stage to annihilate the entries below the kth pivot, it now follows from (3.10.4) and (3.10.5) that

$$\mathbf{A} = \mathbf{LU},$$

where

$$\mathbf{L} = \mathbf{I} + \mathbf{c}_1\mathbf{e}_1^T + \mathbf{c}_2\mathbf{e}_2^T + \cdots + \mathbf{c}_{n-1}\mathbf{e}_{n-1}^T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix} \tag{3.10.6}$$

is the lower-triangular matrix with 1's on the diagonal, and where ℓ_{ij} is precisely the multiplier used to annihilate the (i,j)-position during Gaussian elimination. Thus the factorization $\mathbf{A} = \mathbf{LU}$ can be viewed as the matrix formulation of Gaussian elimination, with the understanding that no row interchanges are used.

LU Factorization

If $\mathbf{A}$ is an $n \times n$ matrix such that a zero pivot is never encountered when applying Gaussian elimination with Type III operations, then $\mathbf{A}$ can be factored as the product $\mathbf{A} = \mathbf{LU}$, where the following hold.

- $\mathbf{L}$ is lower triangular and $\mathbf{U}$ is upper triangular. (3.10.7)
- $\ell_{ii} = 1$ and $u_{ii} \neq 0$ for each $i = 1, 2, \ldots, n$. (3.10.8)
- Below the diagonal of $\mathbf{L}$, the entry ℓ_{ij} is the multiple of row j that is subtracted from row i in order to annihilate the (i,j)-position during Gaussian elimination.
- $\mathbf{U}$ is the final result of Gaussian elimination applied to $\mathbf{A}$.
- The matrices $\mathbf{L}$ and $\mathbf{U}$ are uniquely determined by properties (3.10.7) and (3.10.8).

The decomposition of $\mathbf{A}$ into $\mathbf{A} = \mathbf{LU}$ is called the ***LU factorization of*** $\mathbf{A}$, and the matrices $\mathbf{L}$ and $\mathbf{U}$ are called the ***LU factors of*** $\mathbf{A}$.

Proof. Except for the statement concerning the uniqueness of the LU factors, each point has already been established. To prove uniqueness, observe that LU factors must be nonsingular because they have nonzero diagonals. If $\mathbf{L}_1\mathbf{U}_1 = \mathbf{A} = \mathbf{L}_2\mathbf{U}_2$ are two LU factorizations for $\mathbf{A}$, then

$$\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{U}_2\mathbf{U}_1^{-1}. \tag{3.10.9}$$

Notice that $\mathbf{L}_2^{-1}\mathbf{L}_1$ is lower triangular, while $\mathbf{U}_2\mathbf{U}_1^{-1}$ is upper triangular because the inverse of a matrix that is upper (lower) triangular is again upper (lower) triangular, and because the product of two upper (lower) triangular matrices is also upper (lower) triangular. Consequently, (3.10.9) implies $\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{D} = \mathbf{U}_2\mathbf{U}_1^{-1}$ must be a diagonal matrix. However, $[\mathbf{L}_2]_{ii} = 1 = [\mathbf{L}_2^{-1}]_{ii}$, so it must be the case that $\mathbf{L}_2^{-1}\mathbf{L}_1 = \mathbf{I} = \mathbf{U}_2\mathbf{U}_1^{-1}$, and thus $\mathbf{L}_1 = \mathbf{L}_2$ and $\mathbf{U}_1 = \mathbf{U}_2$. ∎

Example 3.10.1

Once $\mathbf{L}$ and $\mathbf{U}$ are known, there is usually no need to manipulate with $\mathbf{A}$. This together with the fact that the multipliers used in Gaussian elimination occur in just the right places in $\mathbf{L}$ means that $\mathbf{A}$ can be successively overwritten with the information in $\mathbf{L}$ and $\mathbf{U}$ as Gaussian elimination evolves. The rule is to store the multiplier ℓ_{ij} in the position it annihilates—namely, the (i,j)-position of the array. For a 3×3 matrix, the result looks like this:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \xrightarrow{\textit{Type III operations}} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ \ell_{21} & u_{22} & u_{23} \\ \ell_{31} & \ell_{32} & u_{33} \end{pmatrix}.$$

For example, generating the LU factorization of

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 7 & 7 \\ 6 & 18 & 22 \end{pmatrix}$$

by successively overwriting a single 3×3 array would evolve as shown below:

$$\begin{pmatrix} 2 & 2 & 2 \\ 4 & 7 & 7 \\ 6 & 18 & 22 \end{pmatrix} \begin{matrix} R_2 - 2R_1 \\ R_3 - 3R_1 \end{matrix} \longrightarrow \begin{pmatrix} 2 & 2 & 2 \\ 2 & 3 & 3 \\ 3 & 12 & 16 \end{pmatrix} R_3 - 4R_2 \longrightarrow \begin{pmatrix} 2 & 2 & 2 \\ 2 & 3 & 3 \\ 3 & 4 & 4 \end{pmatrix}.$$

Thus

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{pmatrix}.$$

This is an important feature in practical computation because it guarantees that an LU factorization requires no more computer memory than that required to store the original matrix $\mathbf{A}$.

Once the LU factors for a nonsingular matrix $\mathbf{A}_{n\times n}$ have been obtained, it's relatively easy to solve a linear system $\mathbf{Ax} = \mathbf{b}$. By rewriting $\mathbf{Ax} = \mathbf{b}$ as

$$\mathbf{L}(\mathbf{Ux}) = \mathbf{b} \quad \text{and setting} \quad \mathbf{y} = \mathbf{Ux},$$

we see that $\mathbf{Ax} = \mathbf{b}$ is equivalent to the two triangular systems

$$\mathbf{Ly} = \mathbf{b} \quad \text{and} \quad \mathbf{Ux} = \mathbf{y}.$$

First, the lower-triangular system $\mathbf{Ly} = \mathbf{b}$ is solved for $\mathbf{y}$ by ***forward substitution.*** That is, if

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix},$$

set

$$y_1 = b_1, \quad y_2 = b_2 - \ell_{21}y_1, \quad y_3 = b_3 - \ell_{31}y_1 - \ell_{32}y_2, \quad \text{etc.}$$

The forward substitution algorithm can be written more concisely as

$$y_1 = b_1 \quad \text{and} \quad y_i = b_i - \sum_{k=1}^{i-1} \ell_{ik}y_k \quad \text{for} \quad i = 2, 3, \ldots, n. \tag{3.10.10}$$

After $\mathbf{y}$ is known, the upper-triangular system $\mathbf{Ux} = \mathbf{y}$ is solved using the standard back substitution procedure by starting with $x_n = y_n/u_{nn}$, and setting

$$x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{k=i+1}^{n} u_{ik}x_k\right) \quad \text{for} \quad i = n-1,\ n-2,\ \ldots, 1. \tag{3.10.11}$$

It can be verified that only n^2 multiplications/divisions and $n^2 - n$ additions/subtractions are required when (3.10.10) and (3.10.11) are used to solve the two triangular systems $\mathbf{Ly} = \mathbf{b}$ and $\mathbf{Ux} = \mathbf{y}$, so it's relatively cheap to solve $\mathbf{Ax} = \mathbf{b}$ once $\mathbf{L}$ and $\mathbf{U}$ are known—recall from §1.2 that these operation counts are about $n^3/3$ when we start from scratch.

If only one system $\mathbf{Ax} = \mathbf{b}$ is to be solved, then there is no significant difference between the technique of reducing the augmented matrix $[\mathbf{A}|\mathbf{b}]$ to a row echelon form and the LU factorization method presented here. However, suppose it becomes necessary to later solve other systems $\mathbf{Ax} = \tilde{\mathbf{b}}$ with the same coefficient matrix but with different right-hand sides, which is frequently the case in applied work. If the LU factors of $\mathbf{A}$ were computed and saved when the original system was solved, then they need not be recomputed, and the solutions to all subsequent systems $\mathbf{Ax} = \tilde{\mathbf{b}}$ are therefore relatively cheap to obtain. That is, the operation counts for each subsequent system are on the order of n^2, whereas these counts would be on the order of $n^3/3$ if we would start from scratch each time.

Summary

- To solve a nonsingular system $\mathbf{Ax} = \mathbf{b}$ using the LU factorization $\mathbf{A} = \mathbf{LU}$, first solve $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$ with the forward substitution algorithm (3.10.10), and then solve $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$ with the back substitution procedure (3.10.11).

- The advantage of this approach is that once the LU factors for $\mathbf{A}$ have been computed, any other linear system $\mathbf{Ax} = \tilde{\mathbf{b}}$ can be solved with only n^2 multiplications/divisions and $n^2 - n$ additions/subtractions.

Example 3.10.2

Problem 1: Use the LU factorization of $\mathbf{A}$ to solve $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 7 & 7 \\ 6 & 18 & 22 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 12 \\ 24 \\ 12 \end{pmatrix}.$$

Problem 2: Suppose that after solving the original system new information is received that changes $\mathbf{b}$ to

$$\tilde{\mathbf{b}} = \begin{pmatrix} 6 \\ 24 \\ 70 \end{pmatrix}.$$

Use the LU factors of $\mathbf{A}$ to solve the updated system $\mathbf{Ax} = \tilde{\mathbf{b}}$.

Solution 1: The LU factors of the coefficient matrix were determined in Example 3.10.1 to be

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{pmatrix}.$$

The strategy is to set $\mathbf{Ux} = \mathbf{y}$ and solve $\mathbf{Ax} = \mathbf{L}(\mathbf{Ux}) = \mathbf{b}$ by solving the two triangular systems

$$\mathbf{Ly} = \mathbf{b} \quad \text{and} \quad \mathbf{Ux} = \mathbf{y}.$$

First solve the lower-triangular system $\mathbf{Ly} = \mathbf{b}$ by using forward substitution:

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 24 \\ 12 \end{pmatrix} \implies \begin{aligned} y_1 &= 12, \\ y_2 &= 24 - 2y_1 = 0, \\ y_3 &= 12 - 3y_1 - 4y_2 = -24. \end{aligned}$$

Now use back substitution to solve the upper-triangular system $\mathbf{Ux} = \mathbf{y}$:

$$\begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 0 \\ -24 \end{pmatrix} \implies \begin{aligned} x_1 &= (12 - 2x_2 - 2x_3)/2 = 6, \\ x_2 &= (0 - 3x_3)/3 = 6, \\ x_3 &= -24/4 = -6. \end{aligned}$$

Solution 2: To solve the updated system $\mathbf{Ax} = \tilde{\mathbf{b}}$, simply repeat the forward and backward substitution steps with $\mathbf{b}$ replaced by $\tilde{\mathbf{b}}$. Solving $\mathbf{Ly} = \tilde{\mathbf{b}}$ with forward substitution gives the following:

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 24 \\ 70 \end{pmatrix} \implies \begin{aligned} y_1 &= 6, \\ y_2 &= 24 - 2y_1 = 12, \\ y_3 &= 70 - 3y_1 - 4y_2 = 4. \end{aligned}$$

Using back substitution to solve $\mathbf{Ux} = \mathbf{y}$ gives the following updated solution:

$$\begin{pmatrix} 2 & 2 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 12 \\ 4 \end{pmatrix} \implies \begin{aligned} x_1 &= (6 - 2x_2 - 2x_3)/2 = -1, \\ x_2 &= (12 - 3x_3)/3 = 3, \\ x_3 &= 4/4 = 1. \end{aligned}$$

Example 3.10.3

Computing $\mathbf{A}^{-1}$. Although matrix inversion is not used for solving $\mathbf{Ax} = \mathbf{b}$, there are a few applications where explicit knowledge of $\mathbf{A}^{-1}$ is desirable.

Problem: Explain how to use the LU factors of a nonsingular matrix $\mathbf{A}_{n\times n}$ to compute $\mathbf{A}^{-1}$ efficiently.

Solution: The strategy is to solve the matrix equation $\mathbf{AX} = \mathbf{I}$. Recall from (3.5.5) that $\mathbf{AA}^{-1} = \mathbf{I}$ implies $\mathbf{A}[\mathbf{A}^{-1}]_{*j} = \mathbf{e}_j$, so the jth column of $\mathbf{A}^{-1}$ is the solution of a system $\mathbf{Ax}_j = \mathbf{e}_j$. Each of these n systems has the same coefficient matrix, so, once the LU factors for $\mathbf{A}$ are known, each system $\mathbf{Ax}_j = \mathbf{LUx}_j = \mathbf{e}_j$ can be solved by the standard two-step process.

(1) Set $\mathbf{y}_j = \mathbf{Ux}_j$, and solve $\mathbf{Ly}_j = \mathbf{e}_j$ for $\mathbf{y}_j$ by forward substitution.

(2) Solve $\mathbf{Ux}_j = \mathbf{y}_j$ for $\mathbf{x}_j = [\mathbf{A}^{-1}]_{*j}$ by back substitution.

This method has at least two advantages: it's efficient, and any code written to solve $\mathbf{Ax} = \mathbf{b}$ can also be used to compute $\mathbf{A}^{-1}$.

Note: A tempting alternate solution might be to use the fact $\mathbf{A}^{-1} = (\mathbf{LU})^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$. But computing $\mathbf{U}^{-1}$ and $\mathbf{L}^{-1}$ explicitly and then multiplying the results is not as computationally efficient as the method just described.

Not all nonsingular matrices possess an LU factorization. For example, there is clearly no nonzero value of u_{11} that will satisfy

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \ell_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix}.$$

The problem here is the zero pivot in the (1,1)-position. Our development of the LU factorization using elementary lower-triangular matrices shows that if no zero pivots emerge, then no row interchanges are necessary, and the LU factorization can indeed be carried to completion. The converse is also true (its proof is left as an exercise), so we can say that *a nonsingular matrix* $\mathbf{A}$ *has an LU factorization if and only if a zero pivot does not emerge during row reduction to upper-triangular form with Type III operations.*

Although it is a bit more theoretical, there is another interesting way to characterize the existence of LU factors. This characterization is given in terms of the ***leading principal submatrices of*** $\mathbf{A}$ that are defined to be those submatrices taken from the upper-left-hand corner of $\mathbf{A}$. That is,

$$\mathbf{A}_1 = \begin{pmatrix} a_{11} \end{pmatrix}, \quad \mathbf{A}_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \ldots, \mathbf{A}_k = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}, \ldots.$$

Existence of LU Factors

Each of the following statements is equivalent to saying that a nonsingular matrix $\mathbf{A}_{n\times n}$ possesses an LU factorization.

- A zero pivot does not emerge during row reduction to upper-triangular form with Type III operations.
- Each leading principal submatrix $\mathbf{A}_k$ is nonsingular. (3.10.12)

Proof. We will prove the statement concerning the leading principal submatrices and leave the proof concerning the nonzero pivots as an exercise. Assume first that $\mathbf{A}$ possesses an LU factorization and partition $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{LU} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{0} & \mathbf{U}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_{11}\mathbf{U}_{11} & * \\ * & * \end{pmatrix},$$

where $\mathbf{L}_{11}$ and $\mathbf{U}_{11}$ are each $k \times k$. Thus $\mathbf{A}_k = \mathbf{L}_{11}\mathbf{U}_{11}$ must be nonsingular because $\mathbf{L}_{11}$ and $\mathbf{U}_{11}$ are each nonsingular—they are triangular with nonzero diagonal entries. Conversely, suppose that each leading principal submatrix in $\mathbf{A}$ is nonsingular. Use induction to prove that each $\mathbf{A}_k$ possesses an LU factorization. For $k = 1$, this statement is clearly true because if $\mathbf{A}_1 = (a_{11})$ is nonsingular, then $\mathbf{A}_1 = (1)(a_{11})$ is its LU factorization. Now assume that $\mathbf{A}_k$ has an LU factorization and show that this together with the nonsingularity condition implies $\mathbf{A}_{k+1}$ must also possess an LU factorization. If $\mathbf{A}_k = \mathbf{L}_k\mathbf{U}_k$ is the LU factorization for $\mathbf{A}_k$, then $\mathbf{A}_k^{-1} = \mathbf{U}_k^{-1}\mathbf{L}_k^{-1}$ so that

$$\mathbf{A}_{k+1} = \begin{pmatrix} \mathbf{A}_k & \mathbf{b} \\ \mathbf{c}^T & \alpha_{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_k & \mathbf{0} \\ \mathbf{c}^T\mathbf{U}_k^{-1} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{U}_k & \mathbf{L}_k^{-1}\mathbf{b} \\ \mathbf{0} & \alpha_{k+1} - \mathbf{c}^T\mathbf{A}_k^{-1}\mathbf{b} \end{pmatrix}, \quad (3.10.13)$$

where $\mathbf{c}^T$ and $\mathbf{b}$ contain the first k components of $\mathbf{A}_{k+1*}$ and $\mathbf{A}_{*k+1}$, respectively. Observe that this is the LU factorization for $\mathbf{A}_{k+1}$ because

$$\mathbf{L}_{k+1} = \begin{pmatrix} \mathbf{L}_k & \mathbf{0} \\ \mathbf{c}^T\mathbf{U}_k^{-1} & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U}_{k+1} = \begin{pmatrix} \mathbf{U}_k & \mathbf{L}_k^{-1}\mathbf{b} \\ \mathbf{0} & \alpha_{k+1} - \mathbf{c}^T\mathbf{A}_k^{-1}\mathbf{b} \end{pmatrix}$$

are lower- and upper-triangular matrices, respectively, and $\mathbf{L}$ has 1's on its diagonal while the diagonal entries of $\mathbf{U}$ are nonzero. The fact that

$$\alpha_{k+1} - \mathbf{c}^T\mathbf{A}_k^{-1}\mathbf{b} \neq 0$$

follows because $\mathbf{A}_{k+1}$ and $\mathbf{L}_{k+1}$ are each nonsingular, so $\mathbf{U}_{k+1} = \mathbf{L}_{k+1}^{-1}\mathbf{A}_{k+1}$ must also be nonsingular. Therefore, the nonsingularity of the leading principal

submatrices implies that each $\mathbf{A}_k$ possesses an LU factorization, and hence $\mathbf{A}_n = \mathbf{A}$ must have an LU factorization. ∎

Up to this point we have avoided dealing with row interchanges because if a row interchange is needed to remove a zero pivot, then no LU factorization is possible. However, we know from the discussion in §1.5 that practical computation necessitates row interchanges in the form of partial pivoting. So even if no zero pivots emerge, it is usually the case that we must still somehow account for row interchanges.

To understand the effects of row interchanges in the framework of an LU decomposition, let $\mathbf{T}_k = \mathbf{I} - \mathbf{c}_k\mathbf{e}_k^T$ be an elementary lower-triangular matrix as described in (3.10.2), and let $\mathbf{E} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ with $\mathbf{u} = \mathbf{e}_{k+i} - \mathbf{e}_{k+j}$ be the Type I elementary interchange matrix associated with an interchange of rows $k+i$ and $k+j$. Notice that $\mathbf{e}_k^T\mathbf{E} = \mathbf{e}_k^T$ because $\mathbf{e}_k^T$ has 0's in positions $k+i$ and $k+j$. This together with the fact that $\mathbf{E}^2 = \mathbf{I}$ guarantees

$$\mathbf{E}\mathbf{T}_k\mathbf{E} = \mathbf{E}^2 - \mathbf{E}\mathbf{c}_k\mathbf{e}_k^T\mathbf{E} = \mathbf{I} - \tilde{\mathbf{c}}_k\mathbf{e}_k^T, \quad \text{where} \quad \tilde{\mathbf{c}}_k = \mathbf{E}\mathbf{c}_k.$$

In other words, the matrix

$$\tilde{\mathbf{T}}_k = \mathbf{E}\mathbf{T}_k\mathbf{E} = \mathbf{I} - \tilde{\mathbf{c}}_k\mathbf{e}_k^T \tag{3.10.14}$$

is also an elementary lower-triangular matrix, and $\tilde{\mathbf{T}}_k$ agrees with $\mathbf{T}_k$ in all positions except that the multipliers μ_{k+i} and μ_{k+j} have traded places. As before, assume we are row reducing an $n \times n$ nonsingular matrix $\mathbf{A}$, but suppose that an interchange of rows $k+i$ and $k+j$ is necessary immediately after the kth stage so that the sequence of left-hand multiplications $\mathbf{E}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_1$ is applied to $\mathbf{A}$. Since $\mathbf{E}^2 = \mathbf{I}$, we may insert $\mathbf{E}^2$ to the right of each $\mathbf{T}$ to obtain

$$\begin{aligned}\mathbf{E}\mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_1 &= \mathbf{E}\mathbf{T}_k\mathbf{E}^2\mathbf{T}_{k-1}\mathbf{E}^2\cdots\mathbf{E}^2\mathbf{T}_1\mathbf{E}^2 \\ &= (\mathbf{E}\mathbf{T}_k\mathbf{E})\,(\mathbf{E}\mathbf{T}_{k-1}\mathbf{E})\cdots(\mathbf{E}\mathbf{T}_1\mathbf{E})\,\mathbf{E} \\ &= \tilde{\mathbf{T}}_k\tilde{\mathbf{T}}_{k-1}\cdots\tilde{\mathbf{T}}_1\mathbf{E}.\end{aligned}$$

In such a manner, the necessary interchange matrices $\mathbf{E}$ can be "factored" to the far-right-hand side, and the matrices $\tilde{\mathbf{T}}$ retain the desirable feature of being elementary lower-triangular matrices. Furthermore, (3.10.14) implies that $\tilde{\mathbf{T}}_k\tilde{\mathbf{T}}_{k-1}\cdots\tilde{\mathbf{T}}_1$ differs from $\mathbf{T}_k\mathbf{T}_{k-1}\cdots\mathbf{T}_1$ only in the sense that the multipliers in rows $k+i$ and $k+j$ have traded places. Therefore, row interchanges in Gaussian elimination can be accounted for by writing $\tilde{\mathbf{T}}_{n-1}\cdots\tilde{\mathbf{T}}_2\tilde{\mathbf{T}}_1\mathbf{P}\mathbf{A} = \mathbf{U}$, where $\mathbf{P}$ is the product of all elementary interchange matrices used during the reduction and where the $\tilde{\mathbf{T}}_k$'s are elementary lower-triangular matrices in which the multipliers have been permuted according to the row interchanges that were implemented. Since all of the $\tilde{\mathbf{T}}_k$'s are elementary lower-triangular matrices, we may proceed along the same lines discussed in (3.10.4)—(3.10.6) to obtain

$$\mathbf{P}\mathbf{A} = \mathbf{L}\mathbf{U}, \quad \text{where} \quad \mathbf{L} = \tilde{\mathbf{T}}_1^{-1}\tilde{\mathbf{T}}_2^{-1}\cdots\tilde{\mathbf{T}}_{n-1}^{-1}. \tag{3.10.15}$$

When row interchanges are allowed, zero pivots can always be avoided when the original matrix $\mathbf{A}$ is nonsingular. Consequently, we may conclude that *for every nonsingular matrix* $\mathbf{A}$, *there exists a* ***permutation matrix*** $\mathbf{P}$ *(a product of elementary interchange matrices) such that* $\mathbf{PA}$ *has an LU factorization.* Furthermore, because of the observation in (3.10.14) concerning how the multipliers in $\mathbf{T}_k$ and $\tilde{\mathbf{T}}_k$ trade places when a row interchange occurs, and because

$$\tilde{\mathbf{T}}_k^{-1} = \left(\mathbf{I} - \tilde{\mathbf{c}}_k \mathbf{e}_k^T\right)^{-1} = \mathbf{I} + \tilde{\mathbf{c}}_k \mathbf{e}_k^T,$$

it is not difficult to see that the same line of reasoning used to arrive at (3.10.6) can be applied to conclude that the multipliers in the matrix $\mathbf{L}$ in (3.10.15) are permuted according to the row interchanges that are executed. More specifically, *if rows* k *and* $k+i$ *are interchanged to create the* k^{th} *pivot, then the multipliers*

$$\begin{pmatrix} \ell_{k1} & \ell_{k2} & \cdots & \ell_{k,k-1} \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} \ell_{k+i,1} & \ell_{k+i,2} & \cdots & \ell_{k+i,k-1} \end{pmatrix}$$

trade places in the formation of $\mathbf{L}$.

This means that we can proceed just as in the case when no interchanges are used and successively overwrite the array originally containing $\mathbf{A}$ with each multiplier replacing the position it annihilates. Whenever a row interchange occurs, the corresponding multipliers will be correctly interchanged as well. The permutation matrix $\mathbf{P}$ is simply the cumulative record of the various interchanges used, and the information in $\mathbf{P}$ is easily accounted for by a simple technique that is illustrated in the following example.

Example 3.10.4

Problem: Use partial pivoting on the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}$$

and determine the LU decomposition $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{P}$ is the associated permutation matrix.

Solution: As explained earlier, the strategy is to successively overwrite the array $\mathbf{A}$ with components from $\mathbf{L}$ and $\mathbf{U}$. For the sake of clarity, the multipliers ℓ_{ij} are shown in boldface type. Adjoin a "permutation counter column" $\mathbf{p}$ that is initially set to the natural order 1,2,3,4. Permuting components of $\mathbf{p}$ as the various row interchanges are executed will accumulate the desired permutation. The matrix $\mathbf{P}$ is obtained by executing the final permutation residing in $\mathbf{p}$ to the rows of an appropriate size identity matrix:

$$[\mathbf{A}|\mathbf{p}] = \left(\begin{array}{cccc|c} 1 & 2 & -3 & 4 & 1 \\ 4 & 8 & 12 & -8 & 2 \\ 2 & 3 & 2 & 1 & 3 \\ -3 & -1 & 1 & -4 & 4 \end{array}\right) \longrightarrow \left(\begin{array}{cccc|c} 4 & 8 & 12 & -8 & 2 \\ 1 & 2 & -3 & 4 & 1 \\ 2 & 3 & 2 & 1 & 3 \\ -3 & -1 & 1 & -4 & 4 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{rrrr|r} 4 & 8 & 12 & -8 & 2 \\ \mathbf{1/4} & 0 & -6 & 6 & 1 \\ \mathbf{1/2} & -1 & -4 & 5 & 3 \\ \mathbf{-3/4} & 5 & 10 & -10 & 4 \end{array}\right) \longrightarrow \left(\begin{array}{rrrr|r} 4 & 8 & 12 & -8 & 2 \\ \mathbf{-3/4} & 5 & 10 & -10 & 4 \\ \mathbf{1/2} & -1 & -4 & 5 & 3 \\ \mathbf{1/4} & 0 & -6 & 6 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{rrrr|r} 4 & 8 & 12 & -8 & 2 \\ \mathbf{-3/4} & 5 & 10 & -10 & 4 \\ \mathbf{1/2} & \mathbf{-1/5} & -2 & 3 & 3 \\ \mathbf{1/4} & \mathbf{0} & -6 & 6 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{rrrr|r} 4 & 8 & 12 & -8 & 2 \\ \mathbf{-3/4} & 5 & 10 & -10 & 4 \\ \mathbf{1/4} & \mathbf{0} & -6 & 6 & 1 \\ \mathbf{1/2} & \mathbf{-1/5} & -2 & 3 & 3 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{rrrr|r} 4 & 8 & 12 & -8 & 2 \\ \mathbf{-3/4} & 5 & 10 & -10 & 4 \\ \mathbf{1/4} & \mathbf{0} & -6 & 6 & 1 \\ \mathbf{1/2} & \mathbf{-1/5} & \mathbf{1/3} & 1 & 3 \end{array}\right).$$

Therefore,

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

It is easy to combine the advantages of partial pivoting with the LU decomposition in order to solve a nonsingular system $\mathbf{Ax} = \mathbf{b}$. Because permutation matrices are nonsingular, the system $\mathbf{Ax} = \mathbf{b}$ is equivalent to

$$\mathbf{PAx} = \mathbf{Pb},$$

and hence we can employ the LU solution techniques discussed earlier to solve this permuted system. That is, if we have already performed the factorization $\mathbf{PA} = \mathbf{LU}$—as illustrated in Example 3.10.4—then we can solve $\mathbf{Ly} = \mathbf{Pb}$ for $\mathbf{y}$ by forward substitution, and then solve $\mathbf{Ux} = \mathbf{y}$ by back substitution.

It should be evident that the permutation matrix $\mathbf{P}$ is not really needed. All that is necessary is knowledge of the LU factors along with the final permutation contained in the permutation counter column $\mathbf{p}$ illustrated in Example 3.10.4. The column $\tilde{\mathbf{b}} = \mathbf{Pb}$ is simply a rearrangement of the components of $\mathbf{b}$ according to the final permutation shown in $\mathbf{p}$. In other words, the strategy is to first permute $\mathbf{b}$ into $\tilde{\mathbf{b}}$ according to the permutation $\mathbf{p}$, and then solve $\mathbf{Ly} = \tilde{\mathbf{b}}$ followed by $\mathbf{Ux} = \mathbf{y}$.

Example 3.10.5

Problem: Use the LU decomposition obtained with partial pivoting to solve the system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 3 \\ 60 \\ 1 \\ 5 \end{pmatrix}.$$

Solution: The LU decomposition with partial pivoting was computed in Example 3.10.4. Permute the components in $\mathbf{b}$ according to the permutation $\mathbf{p} = (2 \quad 4 \quad 1 \quad 3)$, and call the result $\tilde{\mathbf{b}}$. Now solve $\mathbf{Ly} = \tilde{\mathbf{b}}$ by applying forward substitution:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 60 \\ 5 \\ 3 \\ 1 \end{pmatrix} \implies \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 60 \\ 50 \\ -12 \\ -15 \end{pmatrix}.$$

Then solve $\mathbf{Ux} = \mathbf{y}$ by applying back substitution:

$$\begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 60 \\ 50 \\ -12 \\ -15 \end{pmatrix} \implies \mathbf{x} = \begin{pmatrix} 12 \\ 6 \\ -13 \\ -15 \end{pmatrix}.$$

LU Factorization with Row Interchanges

- For each nonsingular matrix $\mathbf{A}$, there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{PA}$ possesses an LU factorization $\mathbf{PA} = \mathbf{LU}$.

- To compute $\mathbf{L}$, $\mathbf{U}$, and $\mathbf{P}$, successively overwrite the array originally containing $\mathbf{A}$. Replace each entry being annihilated with the multiplier used to execute the annihilation. Whenever row interchanges such as those used in partial pivoting are implemented, the multipliers in the array will automatically be interchanged in the correct manner.

- Although the entire permutation matrix $\mathbf{P}$ is rarely called for, it can be constructed by permuting the rows of the identity matrix $\mathbf{I}$ according to the various interchanges used. These interchanges can be accumulated in a "permutation counter column" $\mathbf{p}$ that is initially in natural order $(1, 2, \ldots, n)$—see Example 3.10.4.

- To solve a nonsingular linear system $\mathbf{Ax} = \mathbf{b}$ using the LU decomposition with partial pivoting, permute the components in $\mathbf{b}$ to construct $\tilde{\mathbf{b}}$ according to the sequence of interchanges used—i.e., according to $\mathbf{p}$—and then solve $\mathbf{Ly} = \tilde{\mathbf{b}}$ by forward substitution followed by the solution of $\mathbf{Ux} = \mathbf{y}$ using back substitution.

Example 3.10.6

The LDU factorization. There's some asymmetry in an LU factorization because the lower factor has 1's on its diagonal while the upper factor has a nonunit diagonal. This is easily remedied by factoring the diagonal entries out of the upper factor as shown below:

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} = \begin{pmatrix} u_{11} & 0 & \cdots & 0 \\ 0 & u_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12}/u_{11} & \cdots & u_{1n}/u_{11} \\ 0 & 1 & \cdots & u_{2n}/u_{22} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Setting $\mathbf{D} = \operatorname{diag}(u_{11}, u_{22}, \ldots, u_{nn})$ (the diagonal matrix of pivots) and redefining $\mathbf{U}$ to be the rightmost upper-triangular matrix shown above allows any LU factorization to be written as $\mathbf{A} = \mathbf{LDU}$, where $\mathbf{L}$ and $\mathbf{U}$ are lower- and upper-triangular matrices with 1's on both of their diagonals. This is called the ***LDU factorization*** of $\mathbf{A}$. It is uniquely determined, and when $\mathbf{A}$ is symmetric, the LDU factorization is $\mathbf{A} = \mathbf{LDL}^T$ (Exercise 3.10.9).

Example 3.10.7

The Cholesky Factorization.[22] A symmetric matrix $\mathbf{A}$ possessing an LU factorization in which each pivot is positive is said to be ***positive definite***.

Problem: Prove that $\mathbf{A}$ is positive definite if and only if $\mathbf{A}$ can be uniquely factored as $\mathbf{A} = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R}$ is an upper-triangular matrix with positive diagonal entries. This is known as the *Cholesky factorization* of $\mathbf{A}$, and $\mathbf{R}$ is called the *Cholesky factor* of $\mathbf{A}$.

Solution: If $\mathbf{A}$ is positive definite, then, as pointed out in Example 3.10.6, it has an LDU factorization $\mathbf{A} = \mathbf{LDL}^T$ in which $\mathbf{D} = \operatorname{diag}(p_1, p_2, \ldots, p_n)$ is the diagonal matrix containing the pivots $p_i > 0$. Setting $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{L}^T$ where $\mathbf{D}^{1/2} = \operatorname{diag}\left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_n}\right)$ yields the desired factorization because $\mathbf{A} = \mathbf{LD}^{1/2}\mathbf{D}^{1/2}\mathbf{L}^T = \mathbf{R}^T\mathbf{R}$, and $\mathbf{R}$ is upper triangular with positive diagonal

[22] This is named in honor of the French military officer Major André-Louis Cholesky (1875–1918). Although originally assigned to an artillery branch, Cholesky later became attached to the Geodesic Section of the Geographic Service in France where he became noticed for his extraordinary intelligence and his facility for mathematics. From 1905 to 1909 Cholesky was involved with the problem of adjusting the triangularization grid for France. This was a huge computational task, and there were arguments as to what computational techniques should be employed. It was during this period that Cholesky invented the ingenious procedure for solving a positive definite system of equations that is the basis for the matrix factorization that now bears his name. Unfortunately, Cholesky's mathematical talents were never allowed to flower. In 1914 war broke out, and Cholesky was again placed in an artillery group—but this time as the commander. On August 31, 1918, Major Cholesky was killed in battle. Cholesky never had time to publish his clever computational methods—they were carried forward by word-of-mouth. Issues surrounding the Cholesky factorization have been independently rediscovered several times by people who were unaware of Cholesky, and, in some circles, the Cholesky factorization is known as the *square root method.*

entries. Conversely, if $\mathbf{A} = \mathbf{R}\mathbf{R}^T$, where $\mathbf{R}$ is lower triangular with a positive diagonal, then factoring the diagonal entries out of $\mathbf{R}$ as illustrated in Example 3.10.6 produces $\mathbf{R} = \mathbf{L}\mathbf{D}$, where $\mathbf{L}$ is lower triangular with a unit diagonal and $\mathbf{D}$ is the diagonal matrix whose diagonal entries are the r_{ii}'s. Consequently, $\mathbf{A} = \mathbf{L}\mathbf{D}^2\mathbf{L}^T$ is the LDU factorization for $\mathbf{A}$, and thus the pivots must be positive because they are the diagonal entries in $\mathbf{D}^2$. We have now proven that $\mathbf{A}$ is positive definite if and only if it has a Cholesky factorization. To see why such a factorization is unique, suppose $\mathbf{A} = \mathbf{R}_1\mathbf{R}_1^T = \mathbf{R}_2\mathbf{R}_2^T$, and factor out the diagonal entries as illustrated in Example 3.10.6 to write $\mathbf{R}_1 = \mathbf{L}_1\mathbf{D}_1$ and $\mathbf{R}_2 = \mathbf{L}_2\mathbf{D}_2$, where each $\mathbf{R}_i$ is lower triangular with a unit diagonal and $\mathbf{D}_i$ contains the diagonal of $\mathbf{R}_i$ so that $\mathbf{A} = \mathbf{L}_1\mathbf{D}_1^2\mathbf{L}_1^T = \mathbf{L}_2\mathbf{D}_2^2\mathbf{L}_2^T$. The uniqueness of the LDU factors insures that $\mathbf{L}_1 = \mathbf{L}_2$ and $\mathbf{D}_1 = \mathbf{D}_2$, so $\mathbf{R}_1 = \mathbf{R}_2$. **Note:** More is said about the Cholesky factorization and positive definite matrices on pp. 313, 345, and 559.

Exercises for section 3.10

3.10.1. Let $\mathbf{A} = \begin{pmatrix} 1 & 4 & 5 \\ 4 & 18 & 26 \\ 3 & 16 & 30 \end{pmatrix}$.

(a) Determine the LU factors of $\mathbf{A}$.

(b) Use the LU factors to solve $\mathbf{A}\mathbf{x}_1 = \mathbf{b}_1$ as well as $\mathbf{A}\mathbf{x}_2 = \mathbf{b}_2$, where

$$\mathbf{b}_1 = \begin{pmatrix} 6 \\ 0 \\ -6 \end{pmatrix} \quad \text{and} \quad \mathbf{b}_2 = \begin{pmatrix} 6 \\ 6 \\ 12 \end{pmatrix}.$$

(c) Use the LU factors to determine $\mathbf{A}^{-1}$.

3.10.2. Let $\mathbf{A}$ and $\mathbf{b}$ be the matrices

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 4 & 17 \\ 3 & 6 & -12 & 3 \\ 2 & 3 & -3 & 2 \\ 0 & 2 & -2 & 6 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 17 \\ 3 \\ 3 \\ 4 \end{pmatrix}.$$

(a) Explain why $\mathbf{A}$ does not have an LU factorization.

(b) Use partial pivoting and find the permutation matrix $\mathbf{P}$ as well as the LU factors such that $\mathbf{P}\mathbf{A} = \mathbf{L}\mathbf{U}$.

(c) Use the information in $\mathbf{P}$, $\mathbf{L}$, and $\mathbf{U}$ to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$.

3.10.3. Determine all values of ξ for which $\mathbf{A} = \begin{pmatrix} \xi & 2 & 0 \\ 1 & \xi & 1 \\ 0 & 1 & \xi \end{pmatrix}$ fails to have an LU factorization.

3.10.4. If $\mathbf{A}$ is a nonsingular matrix that possesses an LU factorization, prove that the pivot that emerges after $(k+1)$ stages of standard Gaussian elimination using only Type III operations is given by

$$p_{k+1} = a_{k+1,k+1} - \mathbf{c}^T \mathbf{A}_k^{-1} \mathbf{b},$$

where $\mathbf{A}_k$ and

$$\mathbf{A}_{k+1} = \begin{pmatrix} \mathbf{A}_k & \mathbf{b} \\ \mathbf{c}^T & a_{k+1,k+1} \end{pmatrix}$$

are the leading principal submatrices of orders k and $k+1$, respectively. Use this to deduce that all pivots must be nonzero when an LU factorization for $\mathbf{A}$ exists.

3.10.5. If $\mathbf{A}$ is a matrix that contains only integer entries and all of its pivots are 1, explain why $\mathbf{A}^{-1}$ must also be an integer matrix. **Note:** This fact can be used to construct random integer matrices that possess integer inverses by randomly generating integer matrices $\mathbf{L}$ and $\mathbf{U}$ with unit diagonals and then constructing the product $\mathbf{A} = \mathbf{LU}$.

3.10.6. Consider the tridiagonal matrix $\mathbf{T} = \begin{pmatrix} \beta_1 & \gamma_1 & 0 & 0 \\ \alpha_1 & \beta_2 & \gamma_2 & 0 \\ 0 & \alpha_2 & \beta_3 & \gamma_3 \\ 0 & 0 & \alpha_3 & \beta_4 \end{pmatrix}$.

(a) Assuming that $\mathbf{T}$ possesses an LU factorization, verify that it is given by

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \alpha_1/\pi_1 & 1 & 0 & 0 \\ 0 & \alpha_2/\pi_2 & 1 & 0 \\ 0 & 0 & \alpha_3/\pi_3 & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} \pi_1 & \gamma_1 & 0 & 0 \\ 0 & \pi_2 & \gamma_2 & 0 \\ 0 & 0 & \pi_3 & \gamma_3 \\ 0 & 0 & 0 & \pi_4 \end{pmatrix},$$

where the π_i 's are generated by the recursion formula

$$\pi_1 = \beta_1 \quad \text{and} \quad \pi_{i+1} = \beta_{i+1} - \frac{\alpha_i \gamma_i}{\pi_i}.$$

Note: This holds for tridiagonal matrices of arbitrary size thereby making the LU factors of these matrices very easy to compute.

(b) Apply the recursion formula given above to obtain the LU factorization of

$$\mathbf{T} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

3.10.7. $\mathbf{A}_{n\times n}$ is called a ***band matrix*** if $a_{ij} = 0$ whenever $|i-j| > w$ for some positive integer w, called the ***bandwidth.*** In other words, the nonzero entries of $\mathbf{A}$ are constrained to be in a band of w diagonal lines above and below the main diagonal. For example, tridiagonal matrices have bandwidth one, and diagonal matrices have bandwidth zero. If $\mathbf{A}$ is a nonsingular matrix with bandwidth w, and if $\mathbf{A}$ has an LU factorization $\mathbf{A} = \mathbf{LU}$, then $\mathbf{L}$ inherits the lower band structure of $\mathbf{A}$, and $\mathbf{U}$ inherits the upper band structure in the sense that $\mathbf{L}$ has "lower bandwidth" w, and $\mathbf{U}$ has "upper bandwidth" w. Illustrate why this is true by using a generic 5×5 matrix with a bandwidth of $w = 2$.

3.10.8. (a) Construct an example of a nonsingular symmetric matrix that fails to possess an LU (or LDU) factorization.

(b) Construct an example of a nonsingular symmetric matrix that has an LU factorization but is not positive definite.

3.10.9. (a) Determine the LDU factors for $\mathbf{A} = \begin{pmatrix} 1 & 4 & 5 \\ 4 & 18 & 26 \\ 3 & 16 & 30 \end{pmatrix}$ (this is the same matrix used in Exercise 3.10.1).

(b) Prove that if a matrix has an LDU factorization, then the LDU factors are uniquely determined.

(c) If $\mathbf{A}$ is symmetric and possesses an LDU factorization, explain why it must be given by $\mathbf{A} = \mathbf{LDL}^T$.

3.10.10. Explain why $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 8 & 12 \\ 3 & 12 & 27 \end{pmatrix}$ is positive definite, and then find the Cholesky factor $\mathbf{R}$.

CHAPTER 4

Vector Spaces

4.1 SPACES AND SUBSPACES

After matrix theory became established toward the end of the nineteenth century, it was realized that many mathematical entities that were considered to be quite different from matrices were in fact quite similar. For example, objects such as points in the plane $\Re^2$, points in 3-space $\Re^3$, polynomials, continuous functions, and differentiable functions (to name only a few) were recognized to satisfy the same additive properties and scalar multiplication properties given in §3.2 for matrices. Rather than studying each topic separately, it was reasoned that it is more efficient and productive to study many topics at one time by studying the common properties that they satisfy. This eventually led to the axiomatic definition of a vector space.

A vector space involves four things—two sets $\mathcal{V}$ and $\mathcal{F}$, and two algebraic operations called vector addition and scalar multiplication.

- $\mathcal{V}$ is a nonempty set of objects called ***vectors***. Although $\mathcal{V}$ can be quite general, we will usually consider $\mathcal{V}$ to be a set of n-tuples or a set of matrices.
- $\mathcal{F}$ is a scalar field—for us $\mathcal{F}$ is either the field $\Re$ of real numbers or the field $\mathcal{C}$ of complex numbers.
- Vector addition (denoted by $\mathbf{x}+\mathbf{y}$) is an operation between elements of $\mathcal{V}$.
- Scalar multiplication (denoted by $\alpha\mathbf{x}$) is an operation between elements of $\mathcal{F}$ and $\mathcal{V}$.

The formal definition of a vector space stipulates how these four things relate to each other. In essence, the requirements are that vector addition and scalar multiplication must obey exactly the same properties given in §3.2 for matrices.

Vector Space Definition

The set $\mathcal{V}$ is called a ***vector space over*** $\mathcal{F}$ when the vector addition and scalar multiplication operations satisfy the following properties.

(A1) $\mathbf{x}+\mathbf{y} \in \mathcal{V}$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$. This is called the ***closure property for vector addition.***

(A2) $(\mathbf{x}+\mathbf{y})+\mathbf{z} = \mathbf{x}+(\mathbf{y}+\mathbf{z})$ for every $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$.

(A3) $\mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

(A4) There is an element $\mathbf{0} \in \mathcal{V}$ such that $\mathbf{x}+\mathbf{0} = \mathbf{x}$ for every $\mathbf{x} \in \mathcal{V}$.

(A5) For each $\mathbf{x} \in \mathcal{V}$, there is an element $(-\mathbf{x}) \in \mathcal{V}$ such that $\mathbf{x}+(-\mathbf{x}) = \mathbf{0}$.

(M1) $\alpha\mathbf{x} \in \mathcal{V}$ for all $\alpha \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{V}$. This is the ***closure property for scalar multiplication.***

(M2) $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$ for all $\alpha, \beta \in \mathcal{F}$ and every $\mathbf{x} \in \mathcal{V}$.

(M3) $\alpha(\mathbf{x}+\mathbf{y}) = \alpha\mathbf{x}+\alpha\mathbf{y}$ for every $\alpha \in \mathcal{F}$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

(M4) $(\alpha+\beta)\mathbf{x} = \alpha\mathbf{x}+\beta\mathbf{x}$ for all $\alpha, \beta \in \mathcal{F}$ and every $\mathbf{x} \in \mathcal{V}$.

(M5) $1\mathbf{x} = \mathbf{x}$ for every $\mathbf{x} \in \mathcal{V}$.

A theoretical algebraic treatment of the subject would concentrate on the logical consequences of these defining properties, but the objectives in this text are different, so we will not dwell on the axiomatic development.[23] Neverthe-

[23] The idea of defining a vector space by using a set of abstract axioms was contained in a general theory published in 1844 by Hermann Grassmann (1808–1887), a theologian and philosopher from Stettin, Poland, who was a self-taught mathematician. But Grassmann's work was originally ignored because he tried to construct a highly abstract self-contained theory, independent of the rest of mathematics, containing nonstandard terminology and notation, and he had a tendency to mix mathematics with obscure philosophy. Grassmann published a complete revision of his work in 1862 but with no more success. Only later was it realized that he had formulated the concepts we now refer to as linear dependence, bases, and dimension. The Italian mathematician Giuseppe Peano (1858–1932) was one of the few people who noticed Grassmann's work, and in 1888 Peano published a condensed interpretation of it. In a small chapter at the end, Peano gave an axiomatic definition of a vector space similar to the one above, but this drew little attention outside of a small group in Italy. The current definition is derived from the 1918 work of the German mathematician Hermann Weyl (1885–1955). Even though Weyl's definition is closer to Peano's than to Grassmann's, Weyl did not mention his Italian predecessor, but he did acknowledge Grassmann's "epoch making work." Weyl's success with the idea was due in part to the fact that he thought of vector spaces in terms of geometry, whereas Grassmann and Peano treated them as abstract algebraic structures. As we will see, it's the geometry that's important.

less, it is important to recognize some of the more significant examples and to understand why they are indeed vector spaces.

Example 4.1.1

Because **(A1)**–**(A5)** are generalized versions of the five additive properties of matrix addition, and **(M1)**–**(M5)** are generalizations of the five scalar multiplication properties given in §3.2, we can say that the following hold.

- The set $\Re^{m\times n}$ of $m \times n$ real matrices is a vector space over $\Re$.
- The set $\mathcal{C}^{m\times n}$ of $m \times n$ complex matrices is a vector space over $\mathcal{C}$.

Example 4.1.2

The ***real coordinate spaces***

$$\Re^{1\times n} = \{(\,x_1 \quad x_2 \quad \cdots \quad x_n\,),\ x_i \in \Re\} \qquad \text{and} \qquad \Re^{n\times 1} = \left\{ \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},\ x_i \in \Re \right\}$$

are special cases of the preceding example, and these will be the object of most of our attention. In the context of vector spaces, it usually makes no difference whether a coordinate vector is depicted as a row or as a column. When the row or column distinction is irrelevant, or when it is clear from the context, we will use the common symbol $\Re^n$ to designate a coordinate space. In those cases where it is important to distinguish between rows and columns, we will explicitly write $\Re^{1\times n}$ or $\Re^{n\times 1}$. Similar remarks hold for complex coordinate spaces.

Although the coordinate spaces will be our primary concern, be aware that there are many other types of mathematical structures that are vector spaces—this was the reason for making an abstract definition at the outset. Listed below are a few examples.

Example 4.1.3

With function addition and scalar multiplication defined by

$$(f+g)(x) = f(x) + g(x) \qquad \text{and} \qquad (\alpha f)(x) = \alpha f(x),$$

the following sets are vector spaces over $\Re$:

- The set of functions mapping the interval $[0,1]$ into $\Re$.
- The set of all real-valued continuous functions defined on $[0,1]$.
- The set of real-valued functions that are differentiable on $[0,1]$.
- The set of all polynomials with real coefficients.

Example 4.1.4

Consider the vector space $\Re^2$, and let

$$\mathcal{L} = \{(x, y) \mid y = \alpha x\}$$

be a line through the origin. $\mathcal{L}$ is a subset of $\Re^2$, but $\mathcal{L}$ is a special kind of subset because $\mathcal{L}$ also satisfies the properties **(A1)**–**(A5)** and **(M1)**–**(M5)** that define a vector space. This shows that it is possible for one vector space to properly contain other vector spaces.

Subspaces

Let $\mathcal{S}$ be a nonempty subset of a vector space $\mathcal{V}$ over $\mathcal{F}$ (symbolically, $\mathcal{S} \subseteq \mathcal{V}$). If $\mathcal{S}$ is also a vector space over $\mathcal{F}$ using the same addition and scalar multiplication operations, then $\mathcal{S}$ is said to be a ***subspace*** of $\mathcal{V}$. It's not necessary to check all 10 of the defining conditions in order to determine if a subset is also a subspace—only the closure conditions **(A1)** and **(M1)** need to be considered. That is, a nonempty subset $\mathcal{S}$ of a vector space $\mathcal{V}$ is a subspace of $\mathcal{V}$ if and only if

(A1) $\quad \mathbf{x}, \mathbf{y} \in \mathcal{S} \implies \mathbf{x} + \mathbf{y} \in \mathcal{S}$

and

(M1) $\quad \mathbf{x} \in \mathcal{S} \implies \alpha\mathbf{x} \in \mathcal{S}$ for all $\alpha \in \mathcal{F}$.

Proof. If $\mathcal{S}$ is a subset of $\mathcal{V}$, then $\mathcal{S}$ automatically inherits all of the vector space properties of $\mathcal{V}$ except **(A1)**, **(A4)**, **(A5)**, and **(M1)**. However, **(A1)** together with **(M1)** implies **(A4)** and **(A5)**. To prove this, observe that **(M1)** implies $(-\mathbf{x}) = (-1)\mathbf{x} \in \mathcal{S}$ for all $\mathbf{x} \in \mathcal{S}$ so that **(A5)** holds. Since $\mathbf{x}$ and $(-\mathbf{x})$ are now both in $\mathcal{S}$, **(A1)** insures that $\mathbf{x} + (-\mathbf{x}) \in \mathcal{S}$, and thus $\mathbf{0} \in \mathcal{S}$. ∎

Example 4.1.5

Given a vector space $\mathcal{V}$, the set $\mathcal{Z} = \{\mathbf{0}\}$ containing only the zero vector is a subspace of $\mathcal{V}$ because **(A1)** and **(M1)** are trivially satisfied. Naturally, this subspace is called the ***trivial subspace.***

Vector addition in $\Re^2$ and $\Re^3$ is easily visualized by using the ***parallelogram law,*** which states that for two vectors $\mathbf{u}$ and $\mathbf{v}$, the sum $\mathbf{u} + \mathbf{v}$ is the vector defined by the diagonal of the parallelogram as shown in Figure 4.1.1.

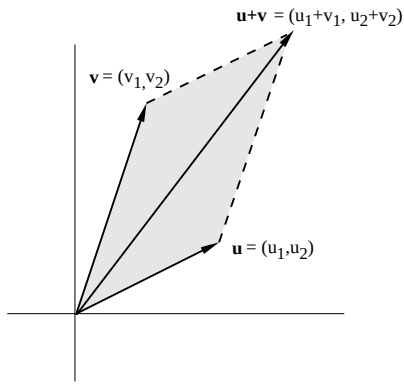


FIGURE 4.1.1

We have already observed that straight lines through the origin in $\Re^2$ are subspaces, but what about straight lines not through the origin? No—they cannot be subspaces because subspaces must contain the zero vector (i.e., they must pass through the origin). What about *curved* lines through the origin—can some of them be subspaces of $\Re^2$? Again the answer is "No!" As depicted in Figure 4.1.2, the parallelogram law indicates why the closure property **(A1)** cannot be satisfied for lines with a curvature because there are points $\mathbf{u}$ and $\mathbf{v}$ on the curve for which $\mathbf{u}+\mathbf{v}$ (the diagonal of the corresponding parallelogram) is not on the curve. Consequently, the only proper subspaces of $\Re^2$ are the trivial subspace and lines through the origin.

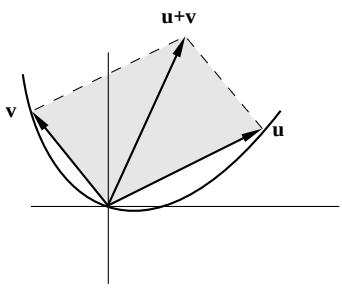


FIGURE 4.1.2

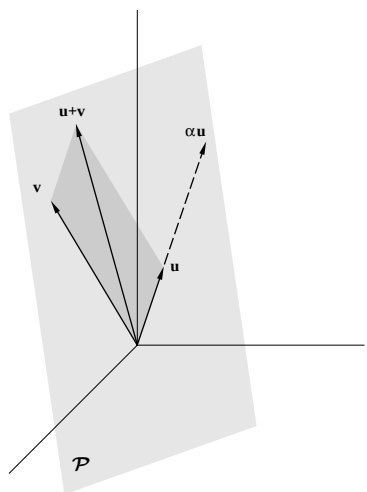


FIGURE 4.1.3

In $\Re^3$, the trivial subspace and lines through the origin are again subspaces, but there is also another one—planes through the origin. If $\mathcal{P}$ is a plane through the origin in $\Re^3$, then, as shown in Figure 4.1.3, the parallelogram law guarantees that the closure property for addition **(A1)** holds—the parallelogram defined by

any two vectors in $\mathcal{P}$ is also in $\mathcal{P}$ so that if $\mathbf{u}, \mathbf{v} \in \mathcal{P}$, then $\mathbf{u} + \mathbf{v} \in \mathcal{P}$. The closure property for scalar multiplication **(M1)** holds because multiplying any vector by a scalar merely stretches it, but its angular orientation does not change so that if $\mathbf{u} \in \mathcal{P}$, then $\alpha\mathbf{u} \in \mathcal{P}$ for all scalars α. Lines and surfaces in $\Re^3$ that have curvature cannot be subspaces for essentially the same reason depicted in Figure 4.1.2. So the only proper subspaces of $\Re^3$ are the trivial subspace, lines through the origin, and planes through the origin.

The concept of a subspace now has an obvious interpretation in the visual spaces $\Re^2$ and $\Re^3$ —*subspaces are the flat surfaces passing through the origin.*

Flatness

Although we can't use our eyes to see "flatness" in higher dimensions, our minds can conceive it through the notion of a subspace. From now on, think of flat surfaces passing through the origin whenever you encounter the term "subspace."

For a set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ from a vector space $\mathcal{V}$, the set of all possible linear combinations of the $\mathbf{v}_i$'s is denoted by

$$span\,(\mathcal{S}) = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_r\mathbf{v}_r \mid \alpha_i \in \mathcal{F}\}\,.$$

Notice that $span\,(\mathcal{S})$ is a subspace of $\mathcal{V}$ because the two closure properties **(A1)** and **(M1)** are satisfied. That is, if $\mathbf{x} = \sum_i \xi_i\mathbf{v}_i$ and $\mathbf{y} = \sum_i \eta_i\mathbf{v}_i$ are two linear combinations from $span\,(\mathcal{S})$, then the sum $\mathbf{x}+\mathbf{y} = \sum_i(\xi_i+\eta_i)\mathbf{v}_i$ is also a linear combination in $span\,(\mathcal{S})$, and for any scalar β, $\beta\mathbf{x} = \sum_i(\beta\xi_i)\mathbf{v}_i$ is also a linear combination in $span\,(\mathcal{S})$.

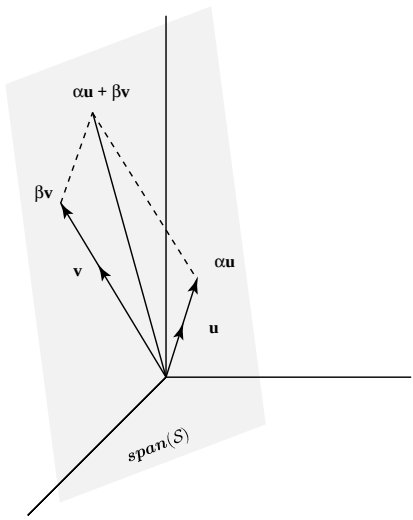


Figure 4.1.4

For example, if $\mathbf{u} \neq \mathbf{0}$ is a vector in $\Re^3$, then $span\{\mathbf{u}\}$ is the straight line passing through the origin and $\mathbf{u}$. If $\mathcal{S} = \{\mathbf{u}, \mathbf{v}\}$, where $\mathbf{u}$ and $\mathbf{v}$ are two nonzero vectors in $\Re^3$ not lying on the same line, then, as shown in Figure 4.1.4, $span(\mathcal{S})$ is the plane passing through the origin and the points $\mathbf{u}$ and $\mathbf{v}$. As we will soon see, *all* subspaces of $\Re^n$ are of the type $span(\mathcal{S})$, so it is worthwhile to introduce the following terminology.

Spanning Sets

- For a set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$, the subspace

$$span(\mathcal{S}) = \{\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_r\mathbf{v}_r\}$$

  generated by forming all linear combinations of vectors from $\mathcal{S}$ is called the ***space spanned by*** $\mathcal{S}$.
- If $\mathcal{V}$ is a vector space such that $\mathcal{V} = span(\mathcal{S})$, we say $\mathcal{S}$ is a ***spanning set*** for $\mathcal{V}$. In other words, $\mathcal{S}$ ***spans*** $\mathcal{V}$ whenever each vector in $\mathcal{V}$ is a linear combination of vectors from $\mathcal{S}$.

Example 4.1.6

(i) In Figure 4.1.4, $\mathcal{S} = \{\mathbf{u}, \mathbf{v}\}$ is a spanning set for the indicated plane.

(ii) $\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\}$ spans the line $y = x$ in $\Re^2$.

(iii) The unit vectors $\left\{ \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$ span $\Re^3$.

(iv) The unit vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ in $\Re^n$ form a spanning set for $\Re^n$.

(v) The finite set $\{1, x, x^2, \ldots, x^n\}$ spans the space of all polynomials such that $\deg p(x) \leq n$, and the infinite set $\{1, x, x^2, \ldots\}$ spans the space of all polynomials.

Example 4.1.7

Problem: For a set of vectors $\mathcal{S} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ from a subspace $\mathcal{V} \subseteq \Re^{m\times 1}$, let $\mathbf{A}$ be the matrix containing the $\mathbf{a}_i$'s as its columns. Explain why $\mathcal{S}$ spans $\mathcal{V}$ if and only if for each $\mathbf{b} \in \mathcal{V}$ there corresponds a column $\mathbf{x}$ such that $\mathbf{Ax} = \mathbf{b}$ (i.e., if and only if $\mathbf{Ax} = \mathbf{b}$ is a consistent system for every $\mathbf{b} \in \mathcal{V}$).

Solution: By definition, $\mathcal{S}$ spans $\mathcal{V}$ if and only if for each $\mathbf{b} \in \mathcal{V}$ there exist scalars α_i such that

$$\mathbf{b} = \alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \cdots + \alpha_n\mathbf{a}_n = \begin{pmatrix} \mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \mathbf{A}\mathbf{x}.$$

Note: This simple observation often is quite helpful. For example, to test whether or not $\mathcal{S} = \{(1 \quad 1 \quad 1), (1 \quad -1 \quad -1), (3 \quad 1 \quad 1)\}$ spans $\Re^3$, place these rows as columns in a matrix $\mathbf{A}$, and ask, "Is the system

$$\begin{pmatrix} 1 & 1 & 3 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

consistent for *every* $\mathbf{b} \in \Re^3$?" Recall from (2.3.4) that $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent if and only if $rank[\mathbf{A}|\mathbf{b}] = rank(\mathbf{A})$. In this case, $rank(\mathbf{A}) = 2$, but $rank[\mathbf{A}|\mathbf{b}] = 3$ for some $\mathbf{b}$'s (e.g., $b_1 = 0$, $b_2 = 1$, $b_3 = 0$), so $\mathcal{S}$ doesn't span $\Re^3$. On the other hand, $\mathcal{S}' = \{(1 \quad 1 \quad 1), (1 \quad -1 \quad -1), (3 \quad 1 \quad 2)\}$ is a spanning set for $\Re^3$ because

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 1 & -1 & 1 \\ 1 & -1 & 2 \end{pmatrix}$$

is nonsingular, so $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent for all $\mathbf{b}$ (the solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$).

As shown below, it's possible to "add" two subspaces to generate another.

Sum of Subspaces

If $\mathcal{X}$ and $\mathcal{Y}$ are subspaces of a vector space $\mathcal{V}$, then the ***sum*** of $\mathcal{X}$ and $\mathcal{Y}$ is defined to be the set of all possible sums of vectors from $\mathcal{X}$ with vectors from $\mathcal{Y}$. That is,

$$\mathcal{X} + \mathcal{Y} = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y}\}.$$

- The sum $\mathcal{X} + \mathcal{Y}$ is again a subspace of $\mathcal{V}$. (4.1.1)
- If $\mathcal{S}_X$, $\mathcal{S}_Y$ span $\mathcal{X}$, $\mathcal{Y}$, then $\mathcal{S}_X \cup \mathcal{S}_Y$ spans $\mathcal{X} + \mathcal{Y}$. (4.1.2)

Proof. To prove (4.1.1), demonstrate that the two closure properties **(A1)** and **(M1)** hold for $\mathcal{S} = \mathcal{X} + \mathcal{Y}$. To show **(A1)** is valid, observe that if $\mathbf{u}, \mathbf{v} \in \mathcal{S}$, then $\mathbf{u} = \mathbf{x}_1 + \mathbf{y}_1$ and $\mathbf{v} = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$. Because $\mathcal{X}$ and $\mathcal{Y}$ are closed with respect to addition, it follows that $\mathbf{x}_1 + \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1 + \mathbf{y}_2 \in \mathcal{Y}$, and therefore $\mathbf{u} + \mathbf{v} = (\mathbf{x}_1 + \mathbf{x}_2) + (\mathbf{y}_1 + \mathbf{y}_2) \in \mathcal{S}$. To verify **(M1)**, observe that $\mathcal{X}$ and $\mathcal{Y}$ are both closed with respect to scalar multiplication so that $\alpha\mathbf{x}_1 \in \mathcal{X}$ and $\alpha\mathbf{y}_1 \in \mathcal{Y}$ for all α, and consequently $\alpha\mathbf{u} = \alpha\mathbf{x}_1 + \alpha\mathbf{y}_1 \in \mathcal{S}$ for all α. To prove (4.1.2), suppose $\mathcal{S}_X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ and $\mathcal{S}_Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\}$, and write

$$\mathbf{z} \in span\,(\mathcal{S}_X \cup \mathcal{S}_Y) \Longleftrightarrow \mathbf{z} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i + \sum_{i=1}^{t} \beta_i \mathbf{y}_i = \mathbf{x} + \mathbf{y} \text{ with } \mathbf{x} \in \mathcal{X},\ \mathbf{y} \in \mathcal{Y}$$
$$\Longleftrightarrow \mathbf{z} \in \mathcal{X} + \mathcal{Y}. \quad \blacksquare$$

Example 4.1.8

If $\mathcal{X} \subseteq \Re^2$ and $\mathcal{Y} \subseteq \Re^2$ are subspaces defined by two different lines through the origin, then $\mathcal{X} + \mathcal{Y} = \Re^2$. This follows from the parallelogram law—sketch a picture for yourself.

Exercises for section 4.1

4.1.1. Determine which of the following subsets of $\Re^n$ are in fact subspaces of $\Re^n$ $(n > 2)$.

(a) $\{\mathbf{x} \mid x_i \geq 0\}$, (b) $\{\mathbf{x} \mid x_1 = 0\}$, (c) $\{\mathbf{x} \mid x_1 x_2 = 0\}$,

(d) $\left\{\mathbf{x} \;\middle|\; \sum_{j=1}^{n} x_j = 0\right\}$, (e) $\left\{\mathbf{x} \;\middle|\; \sum_{j=1}^{n} x_j = 1\right\}$,

(f) $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b},\ \text{where } \mathbf{A}_{m\times n} \neq \mathbf{0} \text{ and } \mathbf{b}_{m\times 1} \neq \mathbf{0}\}$.

4.1.2. Determine which of the following subsets of $\Re^{n\times n}$ are in fact subspaces of $\Re^{n\times n}$.

(a) The symmetric matrices. (b) The diagonal matrices.
(c) The nonsingular matrices. (d) The singular matrices.
(e) The triangular matrices. (f) The upper-triangular matrices.
(g) All matrices that commute with a given matrix $\mathbf{A}$.
(h) All matrices such that $\mathbf{A}^2 = \mathbf{A}$.
(i) All matrices such that $trace\,(\mathbf{A}) = 0$.

4.1.3. If $\mathcal{X}$ is a plane passing through the origin in $\Re^3$ and $\mathcal{Y}$ is the line through the origin that is perpendicular to $\mathcal{X}$, what is $\mathcal{X} + \mathcal{Y}$?

4.1.4. Why must a real or complex nonzero vector space contain an infinite number of vectors?

4.1.5. Sketch a picture in $\Re^3$ of the subspace spanned by each of the following.

(a) $\left\{ \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 6 \\ 4 \end{pmatrix}, \begin{pmatrix} -3 \\ -9 \\ -6 \end{pmatrix} \right\}$, (b) $\left\{ \begin{pmatrix} -4 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}$,

(c) $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$.

4.1.6. Which of the following are spanning sets for $\Re^3$?

(a) $\{(1 \quad 1 \quad 1)\}$ (b) $\{(1 \quad 0 \quad 0), (0 \quad 0 \quad 1)\}$,

(c) $\{(1 \quad 0 \quad 0), (0 \quad 1 \quad 0), (0 \quad 0 \quad 1), (1 \quad 1 \quad 1)\}$,

(d) $\{(1 \quad 2 \quad 1), (2 \quad 0 \quad -1), (4 \quad 4 \quad 1)\}$,

(e) $\{(1 \quad 2 \quad 1), (2 \quad 0 \quad -1), (4 \quad 4 \quad 0)\}$.

4.1.7. For a vector space $\mathcal{V}$, and for $\mathcal{M}, \mathcal{N} \subseteq \mathcal{V}$, explain why $span\,(\mathcal{M} \cup \mathcal{N}) = span\,(\mathcal{M}) + span\,(\mathcal{N})$.

4.1.8. Let $\mathcal{X}$ and $\mathcal{Y}$ be two subspaces of a vector space $\mathcal{V}$.

(a) Prove that the intersection $\mathcal{X} \cap \mathcal{Y}$ is also a subspace of $\mathcal{V}$.

(b) Show that the union $\mathcal{X} \cup \mathcal{Y}$ need not be a subspace of $\mathcal{V}$.

4.1.9. For $\mathbf{A} \in \Re^{m \times n}$ and $\mathcal{S} \subseteq \Re^{n \times 1}$, the set $\mathbf{A}(\mathcal{S}) = \{\mathbf{Ax} \,|\, \mathbf{x} \in \mathcal{S}\}$ contains all possible products of $\mathbf{A}$ with vectors from $\mathcal{S}$. We refer to $\mathbf{A}(\mathcal{S})$ as the set of ***images*** of $\mathcal{S}$ under $\mathbf{A}$.

(a) If $\mathcal{S}$ is a subspace of $\Re^n$, prove $\mathbf{A}(\mathcal{S})$ is a subspace of $\Re^m$.

(b) If $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k$ spans $\mathcal{S}$, show $\mathbf{As}_1, \mathbf{As}_2, \ldots, \mathbf{As}_k$ spans $\mathbf{A}(\mathcal{S})$.

4.1.10. With the usual addition and multiplication, determine whether or not the following sets are vector spaces over the real numbers.

(a) $\Re$, (b) $\mathcal{C}$, (c) The rational numbers.

4.1.11. Let $\mathcal{M} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_r\}$ and $\mathcal{N} = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_r, \mathbf{v}\}$ be two sets of vectors from the same vector space. Prove that $span\,(\mathcal{M}) = span\,(\mathcal{N})$ if and only if $\mathbf{v} \in span\,(\mathcal{M})$.

4.1.12. For a set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, prove that $span\,(\mathcal{S})$ is the intersection of all subspaces that contain $\mathcal{S}$. **Hint:** For $\mathcal{M} = \bigcap\limits_{\mathcal{S} \subseteq \mathcal{V}} \mathcal{V}$, prove that $span\,(\mathcal{S}) \subseteq \mathcal{M}$ and $\mathcal{M} \subseteq span\,(\mathcal{S})$.

4.2 FOUR FUNDAMENTAL SUBSPACES

The closure properties **(A1)** and **(M1)** on p. 162 that characterize the notion of a subspace have much the same "feel" as the definition of a linear function as stated on p. 89, but there's more to it than just a "similar feel." Subspaces are intimately related to linear functions as explained below.

Subspaces and Linear Functions

For a linear function f mapping $\Re^n$ into $\Re^m$, let $\mathcal{R}(f)$ denote the ***range*** of f. That is, $\mathcal{R}(f) = \{f(\mathbf{x}) \,|\, \mathbf{x} \in \Re^n\} \subseteq \Re^m$ is the set of all "images" as $\mathbf{x}$ varies freely over $\Re^n$.

- The range of every linear function $f : \Re^n \to \Re^m$ is a subspace of $\Re^m$, and every subspace of $\Re^m$ is the range of some linear function.

For this reason, subspaces of $\Re^m$ are sometimes called ***linear spaces***.

Proof. If $f : \Re^n \to \Re^m$ is a linear function, then the range of f is a subspace of $\Re^m$ because the closure properties **(A1)** and **(M1)** are satisfied. Establish **(A1)** by showing that $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{R}(f) \Rightarrow \mathbf{y}_1 + \mathbf{y}_2 \in \mathcal{R}(f)$. If $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{R}(f)$, then there must be vectors $\mathbf{x}_1, \mathbf{x}_2 \in \Re^n$ such that $\mathbf{y}_1 = f(\mathbf{x}_1)$ and $\mathbf{y}_2 = f(\mathbf{x}_2)$, so it follows from the linearity of f that

$$\mathbf{y}_1 + \mathbf{y}_2 = f(\mathbf{x}_1) + f(\mathbf{x}_2) = f(\mathbf{x}_1 + \mathbf{x}_2) \in \mathcal{R}(f).$$

Similarly, establish **(M1)** by showing that if $\mathbf{y} \in \mathcal{R}(f)$, then $\alpha\mathbf{y} \in \mathcal{R}(f)$ for all scalars α by using the definition of range along with the linearity of f to write

$$\mathbf{y} \in \mathcal{R}(f) \Longrightarrow \mathbf{y} = f(\mathbf{x}) \text{ for some } \mathbf{x} \in \Re^n \Longrightarrow \alpha\mathbf{y} = \alpha f(\mathbf{x}) = f(\alpha\mathbf{x}) \in \mathcal{R}(f).$$

Now prove that every subspace $\mathcal{V}$ of $\Re^m$ is the range of some linear function $f : \Re^n \to \Re^m$. Suppose that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is a spanning set for $\mathcal{V}$ so that

$$\mathcal{V} = \{\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n \,|\, \alpha_i \in \mathcal{R}\}. \tag{4.2.1}$$

Stack the $\mathbf{v}_i$'s as columns in a matrix $\mathbf{A}_{m\times n} = \big(\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\big)$, and put the α_i's in an $n \times 1$ column $\mathbf{x} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$ to write

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n = \big(\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\big) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \mathbf{A}\mathbf{x}. \tag{4.2.2}$$

The function $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is linear (recall Example 3.6.1, p. 106), and we have that $\mathcal{R}(f) = \{\mathbf{A}\mathbf{x} \,|\, \mathbf{x} \in \Re^{n\times 1}\} = \{\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n \,|\, \alpha_i \in \mathcal{R}\} = \mathcal{V}$. ∎

In particular, this result means that every matrix $\mathbf{A} \in \Re^{m\times n}$ generates a subspace of $\Re^m$ by means of the range of the linear function $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$. Likewise, the transpose[24] of $\mathbf{A} \in \Re^{m\times n}$ defines a subspace of $\Re^n$ by means of the range of $f(\mathbf{y}) = \mathbf{A}^T\mathbf{y}$. These two "range spaces" are two of the four fundamental subspaces defined by a matrix.

Range Spaces

The ***range of a matrix*** $\mathbf{A} \in \Re^{m\times n}$ is defined to be the subspace $R(\mathbf{A})$ of $\Re^m$ that is generated by the range of $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$. That is,

$$R(\mathbf{A}) = \{\mathbf{A}\mathbf{x} \,|\, \mathbf{x} \in \Re^n\} \subseteq \Re^m.$$

Similarly, the range of $\mathbf{A}^T$ is the subspace of $\Re^n$ defined by

$$R\left(\mathbf{A}^T\right) = \{\mathbf{A}^T\mathbf{y} \,|\, \mathbf{y} \in \Re^m\} \subseteq \Re^n.$$

Because $R(\mathbf{A})$ is the set of all "images" of vectors $\mathbf{x} \in \Re^m$ under transformation by $\mathbf{A}$, some people call $R(\mathbf{A})$ the ***image space*** of $\mathbf{A}$.

The observation (4.2.2) that every matrix–vector product $\mathbf{A}\mathbf{x}$ (i.e., every image) is a linear combination of the columns of $\mathbf{A}$ provides a useful characterization of the range spaces. Allowing the components of $\mathbf{x} = (\xi_1, \xi_2, \ldots, \xi_n)^T$ to vary freely and writing

$$\mathbf{A}\mathbf{x} = \left(\mathbf{A}_{*1} \mid \mathbf{A}_{*2} \mid \cdots \mid \mathbf{A}_{*n}\right)\begin{pmatrix}\xi_1\\ \xi_2\\ \vdots\\ \xi_n\end{pmatrix} = \sum_{j=1}^{n} \xi_j \mathbf{A}_{*j}$$

shows that the set of all images $\mathbf{A}\mathbf{x}$ is the same as the set of all linear combinations of the columns of $\mathbf{A}$. Therefore, $R(\mathbf{A})$ is nothing more than the space spanned by the columns of $\mathbf{A}$. That's why $R(\mathbf{A})$ is often called the ***column space*** of $\mathbf{A}$.

Likewise, $R\left(\mathbf{A}^T\right)$ is the space spanned by the columns of $\mathbf{A}^T$. But the columns of $\mathbf{A}^T$ are just the rows of $\mathbf{A}$ (stacked upright), so $R\left(\mathbf{A}^T\right)$ is simply the space spanned by the rows[25] of $\mathbf{A}$. Consequently, $R\left(\mathbf{A}^T\right)$ is also known as the ***row space*** of $\mathbf{A}$. Below is a summary.

---

[24] For ease of exposition, the discussion in this section is in terms of real matrices and real spaces, but all results have complex analogs obtained by replacing $\mathbf{A}^T$ by $\mathbf{A}^*$.

[25] Strictly speaking, the range of $\mathbf{A}^T$ is a set of columns, while the row space of $\mathbf{A}$ is a set of rows. However, no logical difficulties are encountered by considering them to be the same.

Column and Row Spaces

For $\mathbf{A} \in \Re^{m\times n}$, the following statements are true.

- $R(\mathbf{A}) =$ the space spanned by the columns of $\mathbf{A}$ (column space).
- $R\left(\mathbf{A}^T\right) =$ the space spanned by the rows of $\mathbf{A}$ (row space).
- $\mathbf{b} \in R(\mathbf{A}) \Longleftrightarrow \mathbf{b} = \mathbf{A}\mathbf{x}$ for some $\mathbf{x}$. (4.2.3)
- $\mathbf{a} \in R\left(\mathbf{A}^T\right) \Longleftrightarrow \mathbf{a}^T = \mathbf{y}^T\mathbf{A}$ for some $\mathbf{y}^T$. (4.2.4)

Example 4.2.1

Problem: Describe $R(\mathbf{A})$ and $R\left(\mathbf{A}^T\right)$ for $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$.

Solution: $R(\mathbf{A}) = span\{\mathbf{A}_{*1}, \mathbf{A}_{*2}, \mathbf{A}_{*3}\} = \{\alpha_1\mathbf{A}_{*1}+\alpha_2\mathbf{A}_{*2}+\alpha_3\mathbf{A}_{*3} \mid \alpha_i \in \Re\}$, but since $\mathbf{A}_{*2} = 2\mathbf{A}_{*1}$ and $\mathbf{A}_{*3} = 3\mathbf{A}_{*1}$, it's clear that every linear combination of $\mathbf{A}_{*1}$, $\mathbf{A}_{*2}$, and $\mathbf{A}_{*3}$ reduces to a multiple of $\mathbf{A}_{*1}$, so $R(\mathbf{A}) = span\{\mathbf{A}_{*1}\}$. Geometrically, $R(\mathbf{A})$ is the line in $\Re^2$ through the origin and the point $(1, 2)$. Similarly, $R\left(\mathbf{A}^T\right) = span\{\mathbf{A}_{1*}, \mathbf{A}_{2*}\} = \{\alpha_1\mathbf{A}_{1*} + \alpha_2\mathbf{A}_{2*} \mid \alpha_1, \alpha_2 \in \Re\}$. But $\mathbf{A}_{2*} = 2\mathbf{A}_{1*}$ implies that every combination of $\mathbf{A}_{1*}$ and $\mathbf{A}_{2*}$ reduces to a multiple of $\mathbf{A}_{1*}$, so $R\left(\mathbf{A}^T\right) = span\{\mathbf{A}_{1*}\}$, and this is a line in $\Re^3$ through the origin and the point $(1, 2, 3)$.

There are times when it is desirable to know whether or not two matrices have the same row space or the same range. The following theorem provides the solution to this problem.

Equal Ranges

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same shape:

- $R\left(\mathbf{A}^T\right) = R\left(\mathbf{B}^T\right)$ if and only if $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$. (4.2.5)
- $R(\mathbf{A}) = R(\mathbf{B})$ if and only if $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$. (4.2.6)

Proof. To prove (4.2.5), first assume $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$ so that there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{B}$. To see that $R\left(\mathbf{A}^T\right) = R\left(\mathbf{B}^T\right)$, use (4.2.4) to write

$$\begin{aligned}
\mathbf{a} \in R\left(\mathbf{A}^T\right) &\Longleftrightarrow \mathbf{a}^T = \mathbf{y}^T\mathbf{A} = \mathbf{y}^T\mathbf{P}^{-1}\mathbf{PA} \quad \text{for some } \mathbf{y}^T \\
&\Longleftrightarrow \mathbf{a}^T = \mathbf{z}^T\mathbf{B} \quad \text{for } \mathbf{z}^T = \mathbf{y}^T\mathbf{P}^{-1} \\
&\Longleftrightarrow \mathbf{a} \in R\left(\mathbf{B}^T\right).
\end{aligned}$$

Conversely, if $R\left(\mathbf{A}^T\right) = R\left(\mathbf{B}^T\right)$, then

$$span\left\{\mathbf{A}_{1*}, \mathbf{A}_{2*}, \ldots, \mathbf{A}_{m*}\right\} = span\left\{\mathbf{B}_{1*}, \mathbf{B}_{2*}, \ldots, \mathbf{B}_{m*}\right\},$$

so each row of $\mathbf{B}$ is a combination of the rows of $\mathbf{A}$, and vice versa. On the basis of this fact, it can be argued that it is possible to reduce $\mathbf{A}$ to $\mathbf{B}$ by using only row operations (the tedious details are omitted), and thus $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$. The proof of (4.2.6) follows by replacing $\mathbf{A}$ and $\mathbf{B}$ with $\mathbf{A}^T$ and $\mathbf{B}^T$. ∎

Example 4.2.2

Testing Spanning Sets. Two sets $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r\}$ and $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_s\}$ in $\Re^n$ span the same subspace if and only if the nonzero rows of $\mathbf{E_A}$ agree with the nonzero rows of $\mathbf{E_B}$, where $\mathbf{A}$ and $\mathbf{B}$ are the matrices containing the $\mathbf{a}_i$'s and $\mathbf{b}_i$'s as *rows*. This is a corollary of (4.2.5) because zero rows are irrelevant in considering the row space of a matrix, and we already know from (3.9.9) that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$ if and only if $\mathbf{E_A} = \mathbf{E_B}$.

Problem: Determine whether or not the following sets span the same subspace:

$$\mathcal{A} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \\ 1 \\ 4 \end{pmatrix} \right\}, \qquad \mathcal{B} = \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \right\}.$$

Solution: Place the vectors as *rows* in matrices $\mathbf{A}$ and $\mathbf{B}$, and compute

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 2 & 3 \\ 2 & 4 & 1 & 3 \\ 3 & 6 & 1 & 4 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E_A}$$

and

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \to \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} = \mathbf{E_B}.$$

Hence $span\{\mathcal{A}\} = span\{\mathcal{B}\}$ because the nonzero rows in $\mathbf{E_A}$ and $\mathbf{E_B}$ agree.

We already know that the rows of $\mathbf{A}$ span $R\left(\mathbf{A}^T\right)$, and the columns of $\mathbf{A}$ span $R(\mathbf{A})$, but it's often possible to span these spaces with fewer vectors than the full set of rows and columns.

Spanning the Row Space and Range

Let $\mathbf{A}$ be an $m \times n$ matrix, and let $\mathbf{U}$ be any row echelon form derived from $\mathbf{A}$. Spanning sets for the row and column spaces are as follows:

- The nonzero rows of $\mathbf{U}$ span $R\left(\mathbf{A}^T\right)$. (4.2.7)
- The basic columns in $\mathbf{A}$ span $R(\mathbf{A})$. (4.2.8)

Proof. Statement (4.2.7) is an immediate consequence of (4.2.5). To prove (4.2.8), suppose that the basic columns in $\mathbf{A}$ are in positions $b_1, b_2, \ldots, b_r$, and the nonbasic columns occupy positions $n_1, n_2, \ldots, n_t$, and let $\mathbf{Q}_1$ be the permutation matrix that permutes all of the basic columns in $\mathbf{A}$ to the left-hand side so that $\mathbf{AQ}_1 = (\,\mathbf{B}_{m\times r} \quad \mathbf{N}_{m\times t}\,)$, where $\mathbf{B}$ contains the basic columns and $\mathbf{N}$ contains the nonbasic columns. Since the nonbasic columns are linear combinations of the basic columns—recall (2.2.3)—we can annihilate the nonbasic columns in $\mathbf{N}$ using elementary column operations. In other words, there is a nonsingular matrix $\mathbf{Q}_2$ such that $(\,\mathbf{B} \quad \mathbf{N}\,)\,\mathbf{Q}_2 = (\,\mathbf{B} \quad \mathbf{0}\,)$. Thus $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$ is a nonsingular matrix such that $\mathbf{AQ} = \mathbf{AQ}_1\mathbf{Q}_2 = (\,\mathbf{B} \quad \mathbf{N}\,)\,\mathbf{Q}_2 = (\,\mathbf{B} \quad \mathbf{0}\,)$, and hence $\mathbf{A} \overset{\text{col}}{\sim} (\,\mathbf{B} \quad \mathbf{0}\,)$. The conclusion (4.2.8) now follows from (4.2.6). ∎

Example 4.2.3

Problem: Determine spanning sets for $R(\mathbf{A})$ and $R\left(\mathbf{A}^T\right)$, where

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

Solution: Reducing $\mathbf{A}$ to any row echelon form $\mathbf{U}$ provides the solution—the basic columns in $\mathbf{A}$ correspond to the pivotal positions in $\mathbf{U}$, and the nonzero rows of $\mathbf{U}$ span the row space of $\mathbf{A}$. Using $\mathbf{E_A} = \begin{pmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ produces

$$R(\mathbf{A}) = span\left\{\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}\right\} \quad \text{and} \quad R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} 1 \\ 2 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}\right\}.$$

So far, only two of the four fundamental subspaces associated with each matrix $\mathbf{A} \in \Re^{m\times n}$ have been discussed, namely, $R(\mathbf{A})$ and $R\left(\mathbf{A}^T\right)$. To see where the other two fundamental subspaces come from, consider again a general linear function f mapping $\Re^n$ into $\Re^m$, and focus on $\mathcal{N}(f) = \{\mathbf{x} \,|\, f(\mathbf{x}) = \mathbf{0}\}$ (the set of vectors that are mapped to $\mathbf{0}$). $\mathcal{N}(f)$ is called the ***nullspace*** of f (some texts call it the ***kernel*** of f), and it's easy to see that $\mathcal{N}(f)$ is a subspace of $\Re^n$ because the closure properties **(A1)** and **(M1)** are satisfied. Indeed, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{N}(f)$, then $f(\mathbf{x}_1) = \mathbf{0}$ and $f(\mathbf{x}_2) = \mathbf{0}$, so the linearity of f produces

$$f(\mathbf{x}_1 + \mathbf{x}_2) = f(\mathbf{x}_1) + f(\mathbf{x}_2) = \mathbf{0} + \mathbf{0} = \mathbf{0} \implies \mathbf{x}_1 + \mathbf{x}_2 \in \mathcal{N}(f). \quad \textbf{(A1)}$$

Similarly, if $\alpha \in \Re$, and if $\mathbf{x} \in \mathcal{N}(f)$, then $f(\mathbf{x}) = \mathbf{0}$ and linearity implies

$$f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}) = \alpha\mathbf{0} = \mathbf{0} \implies \alpha\mathbf{x} \in \mathcal{N}(f). \quad \textbf{(M1)}$$

By considering the linear functions $f(\mathbf{x}) = \mathbf{Ax}$ and $g(\mathbf{y}) = \mathbf{A}^T\mathbf{y}$, the other two fundamental subspaces defined by $\mathbf{A} \in \Re^{m\times n}$ are obtained. They are $\mathcal{N}(f) = \{\mathbf{x}_{n\times 1} \mid \mathbf{Ax} = \mathbf{0}\} \subseteq \Re^n$ and $\mathcal{N}(g) = \left\{\mathbf{y}_{m\times 1} \mid \mathbf{A^T y} = \mathbf{0}\right\} \subseteq \Re^m$.

Nullspace

- For an $m \times n$ matrix $\mathbf{A}$, the set $N(\mathbf{A}) = \{\mathbf{x}_{n\times 1} \mid \mathbf{A}\mathbf{x} = \mathbf{0}\} \subseteq \Re^n$ is called the ***nullspace*** of $\mathbf{A}$. In other words, $N(\mathbf{A})$ is simply the set of all solutions to the homogeneous system $\mathbf{A}\mathbf{x} = \mathbf{0}$.
- The set $N\left(\mathbf{A}^T\right) = \left\{\mathbf{y}_{m\times 1} \mid \mathbf{A}^T\mathbf{y} = \mathbf{0}\right\} \subseteq \Re^m$ is called the ***left-hand nullspace*** of $\mathbf{A}$ because $N\left(\mathbf{A}^T\right)$ is the set of all solutions to the left-hand homogeneous system $\mathbf{y}^T\mathbf{A} = \mathbf{0}^T$.

Example 4.2.4

Problem: Determine a spanning set for $N(\mathbf{A})$, where $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix}$.

Solution: $N(\mathbf{A})$ is merely the general solution of $\mathbf{A}\mathbf{x} = \mathbf{0}$, and this is determined by reducing $\mathbf{A}$ to a row echelon form $\mathbf{U}$. As discussed in §2.4, any such $\mathbf{U}$ will suffice, so we will use $\mathbf{E}_{\mathbf{A}} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \end{pmatrix}$. Consequently, $x_1 = -2x_2 - 3x_3$, where x_2 and x_3 are free, so the general solution of $\mathbf{A}\mathbf{x} = \mathbf{0}$ is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -2x_2 - 3x_3 \\ x_2 \\ x_3 \end{pmatrix} = x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix}.$$

In other words, $N(\mathbf{A})$ is the set of all possible linear combinations of the vectors

$$\mathbf{h}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{h}_2 = \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix},$$

and therefore $span\{\mathbf{h}_1, \mathbf{h}_2\} = N(\mathbf{A})$. For this example, $N(\mathbf{A})$ is the plane in $\Re^3$ that passes through the origin and the two points $\mathbf{h}_1$ and $\mathbf{h}_2$.

Example 4.2.4 indicates the general technique for determining a spanning set for $N(\mathbf{A})$. Below is a formal statement of this procedure.

Spanning the Nullspace

To determine a spanning set for $N(\mathbf{A})$, where $rank(\mathbf{A}_{m\times n}) = r$, row reduce $\mathbf{A}$ to a row echelon form $\mathbf{U}$, and solve $\mathbf{Ux} = \mathbf{0}$ for the basic variables in terms of the free variables to produce the general solution of $\mathbf{Ax} = \mathbf{0}$ in the form

$$\mathbf{x} = x_{f_1}\mathbf{h}_1 + x_{f_2}\mathbf{h}_2 + \cdots + x_{f_{n-r}}\mathbf{h}_{n-r}. \tag{4.2.9}$$

By definition, the set $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}\}$ spans $N(\mathbf{A})$. Moreover, it can be proven that $\mathcal{H}$ is unique in the sense that $\mathcal{H}$ is independent of the row echelon form $\mathbf{U}$.

It was established in §2.4 that a homogeneous system $\mathbf{Ax} = \mathbf{0}$ possesses a unique solution (i.e., only the trivial solution $\mathbf{x} = \mathbf{0}$) if and only if the rank of the coefficient matrix equals the number of unknowns. This may now be restated using vector space terminology.

Zero Nullspace

If $\mathbf{A}$ is an $m \times n$ matrix, then

- $N(\mathbf{A}) = \{\mathbf{0}\}$ if and only if $rank(\mathbf{A}) = n$; (4.2.10)
- $N(\mathbf{A}^T) = \{\mathbf{0}\}$ if and only if $rank(\mathbf{A}) = m$. (4.2.11)

Proof. We already know that the trivial solution $\mathbf{x} = \mathbf{0}$ is the only solution to $\mathbf{Ax} = \mathbf{0}$ if and only if the rank of $\mathbf{A}$ is the number of unknowns, and this is what (4.2.10) says. Similarly, $\mathbf{A}^T\mathbf{y} = \mathbf{0}$ has only the trivial solution $\mathbf{y} = \mathbf{0}$ if and only if $rank(\mathbf{A}^T) = m$. Recall from (3.9.11) that $rank(\mathbf{A}^T) = rank(\mathbf{A})$ in order to conclude that (4.2.11) holds. ∎

Finally, let's think about how to determine a spanning set for $N(\mathbf{A}^T)$. Of course, we can proceed in the same manner as described in Example 4.2.4 by reducing $\mathbf{A}^T$ to a row echelon form to extract the general solution for $\mathbf{A}^T\mathbf{x} = \mathbf{0}$. However, the other three fundamental subspaces are derivable directly from $\mathbf{E_A}$ (or any other row echelon form $\mathbf{U} \overset{\text{row}}{\sim} \mathbf{A}$), so it's rather awkward to have to start from scratch and compute a new echelon form just to get a spanning set for $N(\mathbf{A}^T)$. It would be better if a single reduction to echelon form could produce all four of the fundamental subspaces. Note that $\mathbf{E}_{\mathbf{A}^T} \neq \mathbf{E}_{\mathbf{A}}^T$, so $\mathbf{E}_{\mathbf{A}}^T$ won't easily lead to $N(\mathbf{A}^T)$. The following theorem helps resolve this issue.

Left-Hand Nullspace

If $rank\,(\mathbf{A}_{m\times n}) = r$, and if $\mathbf{PA} = \mathbf{U}$, where $\mathbf{P}$ is nonsingular and $\mathbf{U}$ is in row echelon form, then the last $m-r$ rows in $\mathbf{P}$ span the left-hand nullspace of $\mathbf{A}$. In other words, if $\mathbf{P} = \begin{pmatrix}\mathbf{P}_1\\ \mathbf{P}_2\end{pmatrix}$, where $\mathbf{P}_2$ is $(m-r)\times m$, then

$$N\left(\mathbf{A}^T\right) = R\left(\mathbf{P}_2^T\right). \tag{4.2.12}$$

Proof. If $\mathbf{U} = \begin{pmatrix}\mathbf{C}\\ \mathbf{0}\end{pmatrix}$, where $\mathbf{C}_{r\times n}$, then $\mathbf{PA} = \mathbf{U}$ implies $\mathbf{P}_2\mathbf{A} = \mathbf{0}$, and this says $R\left(\mathbf{P}_2^T\right) \subseteq N\left(\mathbf{A}^T\right)$. To show equality, demonstrate containment in the opposite direction by arguing that every vector in $N\left(\mathbf{A}^T\right)$ must also be in $R\left(\mathbf{P}_2^T\right)$. Suppose $\mathbf{y}^T \in N\left(\mathbf{A}^T\right)$, and let $\mathbf{P}^{-1} = \begin{pmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{pmatrix}$ to conclude that

$$\mathbf{0} = \mathbf{y}^T\mathbf{A} = \mathbf{y}^T\mathbf{P}^{-1}\mathbf{U} = \mathbf{y}^T\mathbf{Q}_1\mathbf{C} \implies \mathbf{0} = \mathbf{y}^T\mathbf{Q}_1$$

because $N\left(\mathbf{C}^T\right) = \{\mathbf{0}\}$ by (4.2.11). Now observe that $\mathbf{PP}^{-1} = \mathbf{I} = \mathbf{P}^{-1}\mathbf{P}$ insures $\mathbf{P}_1\mathbf{Q}_1 = \mathbf{I}_r$ and $\mathbf{Q}_1\mathbf{P}_1 = \mathbf{I}_m - \mathbf{Q}_2\mathbf{P}_2$, so

$$\begin{aligned}\mathbf{0} = \mathbf{y}^T\mathbf{Q}_1 &\implies \mathbf{0} = \mathbf{y}^T\mathbf{Q}_1\mathbf{P}_1 = \mathbf{y}^T\left(\mathbf{I} - \mathbf{Q}_2\mathbf{P}_2\right)\\ &\implies \mathbf{y}^T = \mathbf{y}^T\mathbf{Q}_2\mathbf{P}_2 = \left(\mathbf{y}^T\mathbf{Q}_2\right)\mathbf{P}_2\\ &\implies \mathbf{y} \in R\left(\mathbf{P}_2^T\right) \implies \mathbf{y}^T \in R\left(\mathbf{P}_2^T\right). \quad \blacksquare\end{aligned}$$

Example 4.2.5

Problem: Determine a spanning set for $N\left(\mathbf{A}^T\right)$, where $\mathbf{A} = \begin{pmatrix}1&2&2&3\\2&4&1&3\\3&6&1&4\end{pmatrix}$.

Solution: To find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{U}$ is in row echelon form, proceed as described in Exercise 3.9.1 and row reduce the augmented matrix $\left(\mathbf{A} \mid \mathbf{I}\right)$ to $\left(\mathbf{U} \mid \mathbf{P}\right)$. It must be the case that $\mathbf{PA} = \mathbf{U}$ because $\mathbf{P}$ is the product of the elementary matrices corresponding to the elementary row operations used. Since any row echelon form will suffice, we may use Gauss–Jordan reduction to reduce $\mathbf{A}$ to $\mathbf{E_A}$ as shown below:

$$\left(\begin{array}{cccc|ccc}1&2&2&3&1&0&0\\2&4&1&3&0&1&0\\3&6&1&4&0&0&1\end{array}\right) \longrightarrow \left(\begin{array}{cccc|ccc}1&2&0&1&-1/3&2/3&0\\0&0&1&1&2/3&-1/3&0\\0&0&0&0&1/3&-5/3&1\end{array}\right)$$

$$\mathbf{P} = \begin{pmatrix}-1/3&2/3&0\\2/3&-1/3&0\\1/3&-5/3&1\end{pmatrix}, \text{ so (4.2.12) implies } N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix}1/3\\-5/3\\1\end{pmatrix}\right\}.$$

Example 4.2.6

Problem: Suppose $rank\,(\mathbf{A}_{m\times n}) = r$, and let $\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{pmatrix}$ be a nonsingular matrix such that $\mathbf{PA} = \mathbf{U} = \begin{pmatrix} \mathbf{C}_{r\times n} \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{U}$ is in row echelon form. Prove

$$R\,(\mathbf{A}) = N\,(\mathbf{P}_2). \tag{4.2.13}$$

Solution: The strategy is to first prove $R\,(\mathbf{A}) \subseteq N\,(\mathbf{P}_2)$ and then show the reverse inclusion $N\,(\mathbf{P}_2) \subseteq R\,(\mathbf{A})$. The equation $\mathbf{PA} = \mathbf{U}$ implies $\mathbf{P}_2\mathbf{A} = \mathbf{0}$, so all columns of $\mathbf{A}$ are in $N\,(\mathbf{P}_2)$, and thus $R\,(\mathbf{A}) \subseteq N\,(\mathbf{P}_2)$. To show inclusion in the opposite direction, suppose $\mathbf{b} \in N\,(\mathbf{P}_2)$, so that

$$\mathbf{Pb} = \begin{pmatrix} \mathbf{P}_1 \\ \mathbf{P}_2 \end{pmatrix} \mathbf{b} = \begin{pmatrix} \mathbf{P}_1\mathbf{b} \\ \mathbf{P}_2\mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{d}_{r\times 1} \\ \mathbf{0} \end{pmatrix}.$$

Consequently, $\mathbf{P}\big(\mathbf{A}\,|\,\mathbf{b}\big) = \big(\mathbf{PA} \mid \mathbf{Pb}\big) = \begin{pmatrix} \mathbf{C} & \mathbf{d} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, and this implies

$$rank[\mathbf{A}|\mathbf{b}] = r = rank\,(\mathbf{A}).$$

Recall from (2.3.4) that this means the system $\mathbf{Ax} = \mathbf{b}$ is consistent, and thus $\mathbf{b} \in R\,(\mathbf{A})$ by (4.2.3). Therefore, $N\,(\mathbf{P}_2) \subseteq R\,(\mathbf{A})$, and we may conclude that $N\,(\mathbf{P}_2) = R\,(\mathbf{A})$.

It's often important to know when two matrices have the same nullspace (or left-hand nullspace). Below is one test for determining this.

Equal Nullspaces

For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same shape:

- $N\,(\mathbf{A}) = N\,(\mathbf{B})$ if and only if $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B}$. (4.2.14)
- $N\left(\mathbf{A}^T\right) = N\left(\mathbf{B}^T\right)$ if and only if $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$. (4.2.15)

Proof. We will prove (4.2.15). If $N\left(\mathbf{A}^T\right) = N\left(\mathbf{B}^T\right)$, then (4.2.12) guarantees $R\left(\mathbf{P}_2^T\right) = N\left(\mathbf{B}^T\right)$, and hence $\mathbf{P}_2\mathbf{B} = \mathbf{0}$. But this means the columns of $\mathbf{B}$ are in $N\,(\mathbf{P}_2)$. That is, $R\,(\mathbf{B}) \subseteq N\,(\mathbf{P}_2) = R\,(\mathbf{A})$ by using (4.2.13). If $\mathbf{A}$ is replaced by $\mathbf{B}$ in the preceding argument—and in (4.2.13)— the result is that $R\,(\mathbf{A}) \subseteq R\,(\mathbf{B})$, and consequently we may conclude that $R\,(\mathbf{A}) = R\,(\mathbf{B})$. The desired conclusion (4.2.15) follows from (4.2.6). Statement (4.2.14) now follows by replacing $\mathbf{A}$ and $\mathbf{B}$ by $\mathbf{A}^T$ and $\mathbf{B}^T$ in (4.2.15). ∎

Summary

The four fundamental subspaces associated with $\mathbf{A}_{m\times n}$ are as follows.

- The range or column space: $R(\mathbf{A}) = \{\mathbf{Ax}\} \subseteq \Re^m$.
- The row space or left-hand range: $R\left(\mathbf{A}^T\right) = \left\{\mathbf{A}^T\mathbf{y}\right\} \subseteq \Re^n$.
- The nullspace: $N(\mathbf{A}) = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{0}\} \subseteq \Re^n$.
- The left-hand nullspace: $N\left(\mathbf{A}^T\right) = \left\{\mathbf{y} \mid \mathbf{A}^T\mathbf{y} = \mathbf{0}\right\} \subseteq \Re^m$.

Let $\mathbf{P}$ be a nonsingular matrix such that $\mathbf{PA} = \mathbf{U}$, where $\mathbf{U}$ is in row echelon form, and suppose $rank(\mathbf{A}) = r$.

- Spanning set for $R(\mathbf{A})$ = the basic columns in $\mathbf{A}$.
- Spanning set for $R\left(\mathbf{A}^T\right)$ = the nonzero rows in $\mathbf{U}$.
- Spanning set for $N(\mathbf{A})$ =the $\mathbf{h}_i$'s in the general solution of $\mathbf{Ax} = \mathbf{0}$.
- Spanning set for $N\left(\mathbf{A}^T\right)$ = the last $m - r$ rows of $\mathbf{P}$.

If $\mathbf{A}$ and $\mathbf{B}$ have the same shape, then

- $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \Longleftrightarrow N(\mathbf{A}) = N(\mathbf{B}) \Longleftrightarrow R\left(\mathbf{A}^T\right) = R\left(\mathbf{B}^T\right)$.
- $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B} \Longleftrightarrow R(\mathbf{A}) = R(\mathbf{B}) \Longleftrightarrow N\left(\mathbf{A}^T\right) = N\left(\mathbf{B}^T\right)$.

Exercises for section 4.2

4.2.1. Determine spanning sets for each of the four fundamental subspaces associated with

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 & 1 & 5 \\ -2 & -4 & 0 & 4 & -2 \\ 1 & 2 & 2 & 4 & 9 \end{pmatrix}.$$

4.2.2. Consider a linear system of equations $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{b}$.

(a) Explain why $\mathbf{Ax} = \mathbf{b}$ is consistent if and only if $\mathbf{b} \in R(\mathbf{A})$.

(b) Explain why a consistent system $\mathbf{Ax} = \mathbf{b}$ has a unique solution if and only if $N(\mathbf{A}) = \{\mathbf{0}\}$.

4.2.3. Suppose that $\mathbf{A}$ is a 3×3 matrix such that

$$\mathcal{R} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} \right\}$$

span $R(\mathbf{A})$ and $N(\mathbf{A})$, respectively, and consider a linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = \begin{pmatrix} 1 \\ -7 \\ 0 \end{pmatrix}$.

(a) Explain why $\mathbf{Ax} = \mathbf{b}$ must be consistent.
(b) Explain why $\mathbf{Ax} = \mathbf{b}$ cannot have a unique solution.

4.2.4. If $\mathbf{A} = \begin{pmatrix} -1 & 1 & 1 & -2 & 1 \\ -1 & 0 & 3 & -4 & 2 \\ -1 & 0 & 3 & -5 & 3 \\ -1 & 0 & 3 & -6 & 4 \\ -1 & 0 & 3 & -6 & 4 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} -2 \\ -5 \\ -6 \\ -7 \\ -7 \end{pmatrix}$, is $\mathbf{b} \in R(\mathbf{A})$?

4.2.5. Suppose that $\mathbf{A}$ is an $n \times n$ matrix.
(a) If $R(\mathbf{A}) = \Re^n$, explain why $\mathbf{A}$ must be nonsingular.
(b) If $\mathbf{A}$ is nonsingular, describe its four fundamental subspaces.

4.2.6. Consider the matrices $\mathbf{A} = \begin{pmatrix} 1 & 1 & 5 \\ 2 & 0 & 6 \\ 1 & 2 & 7 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & -4 & 4 \\ 4 & -8 & 6 \\ 0 & -4 & 5 \end{pmatrix}$.
(a) Do $\mathbf{A}$ and $\mathbf{B}$ have the same row space?
(b) Do $\mathbf{A}$ and $\mathbf{B}$ have the same column space?
(c) Do $\mathbf{A}$ and $\mathbf{B}$ have the same nullspace?
(d) Do $\mathbf{A}$ and $\mathbf{B}$ have the same left-hand nullspace?

4.2.7. If $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$ is a square matrix such that $N(\mathbf{A}_1) = R\left(\mathbf{A}_2^T\right)$, prove that $\mathbf{A}$ must be nonsingular.

4.2.8. Consider a linear system of equations $\mathbf{Ax} = \mathbf{b}$ for which $\mathbf{y}^T\mathbf{b} = 0$ for every $\mathbf{y} \in N\left(\mathbf{A}^T\right)$. Explain why this means the system must be consistent.

4.2.9. For matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{m\times p}$, prove that

$$R(\mathbf{A} \mid \mathbf{B}) = R(\mathbf{A}) + R(\mathbf{B}).$$

4.2.10. Let $\mathbf{p}$ be one particular solution of a linear system $\mathbf{Ax} = \mathbf{b}$.

(a) Explain the significance of the set

$$\mathbf{p} + N(\mathbf{A}) = \{\mathbf{p} + \mathbf{h} \mid \mathbf{h} \in N(\mathbf{A})\}.$$

(b) If $rank(\mathbf{A}_{3\times 3}) = 1$, sketch a picture of $\mathbf{p} + N(\mathbf{A})$ in $\Re^3$.

(c) Repeat part (b) for the case when $rank(\mathbf{A}_{3\times 3}) = 2$.

4.2.11. Suppose that $\mathbf{Ax} = \mathbf{b}$ is a consistent system of linear equations, and let $\mathbf{a} \in R\left(\mathbf{A}^T\right)$. Prove that the inner product $\mathbf{a}^T\mathbf{x}$ is constant for all solutions to $\mathbf{Ax} = \mathbf{b}$.

4.2.12. For matrices such that the product $\mathbf{AB}$ is defined, explain why each of the following statements is true.

(a) $R(\mathbf{AB}) \subseteq R(\mathbf{A})$.

(b) $N(\mathbf{AB}) \supseteq N(\mathbf{B})$.

4.2.13. Suppose that $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ is a spanning set for $R(\mathbf{B})$. Prove that $\mathbf{A}(\mathcal{B}) = \{\mathbf{Ab}_1, \mathbf{Ab}_2, \ldots, \mathbf{Ab}_n\}$ is a spanning set for $R(\mathbf{AB})$.

4.3 LINEAR INDEPENDENCE

For a given set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ there may or may not exist dependency relationships in the sense that it may or may not be possible to express one vector as a linear combination of the others. For example, in the set

$$\mathcal{A} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 9 \\ -3 \\ 4 \end{pmatrix} \right\},$$

the third vector is a linear combination of the first two—i.e., $\mathbf{v}_3 = 3\mathbf{v}_1 + 2\mathbf{v}_2$. Such a dependency always can be expressed in terms of a homogeneous equation by writing

$$3\mathbf{v}_1 + 2\mathbf{v}_2 - \mathbf{v}_3 = \mathbf{0}.$$

On the other hand, it is evident that there are no dependency relationships in the set

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

because no vector can be expressed as a combination of the others. Another way to say this is to state that there are no solutions for α_1, α_2, and α_3 in the homogeneous equation

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 = \mathbf{0}$$

other than the trivial solution $\alpha_1 = \alpha_2 = \alpha_3 = 0$. These observations are the basis for the following definitions.

Linear Independence

A set of vectors $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is said to be a ***linearly independent set*** whenever the only solution for the scalars α_i in the homogeneous equation

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n = \mathbf{0} \tag{4.3.1}$$

is the trivial solution $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Whenever there is a nontrivial solution for the α's (i.e., at least one $\alpha_i \neq 0$) in (4.3.1), the set $\mathcal{S}$ is said to be a ***linearly dependent set.*** In other words, linearly independent sets are those that contain no dependency relationships, and linearly dependent sets are those in which at least one vector is a combination of the others. We will agree that the empty set is always linearly independent.

It is important to realize that the concepts of linear independence and dependence are defined only for sets—individual vectors are neither linearly independent nor dependent. For example consider the following sets:

$$\mathcal{S}_1 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{S}_2 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{S}_3 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

It should be clear that $\mathcal{S}_1$ and $\mathcal{S}_2$ are linearly independent sets while $\mathcal{S}_3$ is linearly dependent. This shows that individual vectors can simultaneously belong to linearly independent sets as well as linearly dependent sets. Consequently, it makes no sense to speak of "linearly independent vectors" or "linearly dependent vectors."

Example 4.3.1

Problem: Determine whether or not the set

$$\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} \right\}$$

is linearly independent.

Solution: Simply determine whether or not there exists a nontrivial solution for the α's in the homogeneous equation

$$\alpha_1 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \alpha_3 \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

or, equivalently, if there is a nontrivial solution to the homogeneous system

$$\begin{pmatrix} 1 & 1 & 5 \\ 2 & 0 & 6 \\ 1 & 2 & 7 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

If $\mathbf{A} = \begin{pmatrix} 1 & 1 & 5 \\ 2 & 0 & 6 \\ 1 & 2 & 7 \end{pmatrix}$, then $\mathbf{E_A} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$, and therefore there exist nontrivial solutions. Consequently, $\mathcal{S}$ is a linearly dependent set. Notice that one particular dependence relationship in $\mathcal{S}$ is revealed by $\mathbf{E_A}$ because it guarantees that $\mathbf{A}_{*3} = 3\mathbf{A}_{*1} + 2\mathbf{A}_{*2}$. This example indicates why the question of whether or not a subset of $\Re^m$ is linearly independent is really a question about whether or not the nullspace of an associated matrix is trivial. The following is a more formal statement of this fact.

Linear Independence and Matrices

Let $\mathbf{A}$ be an $m \times n$ matrix.

- Each of the following statements is equivalent to saying that the columns of $\mathbf{A}$ form a linearly independent set.
  - ▷ $N(\mathbf{A}) = \{\mathbf{0}\}$. (4.3.2)
  - ▷ $rank(\mathbf{A}) = n$. (4.3.3)
- Each of the following statements is equivalent to saying that the rows of $\mathbf{A}$ form a linearly independent set.
  - ▷ $N(\mathbf{A}^T) = \{\mathbf{0}\}$. (4.3.4)
  - ▷ $rank(\mathbf{A}) = m$. (4.3.5)
- When $\mathbf{A}$ is a square matrix, each of the following statements is equivalent to saying that $\mathbf{A}$ is nonsingular.
  - ▷ The columns of $\mathbf{A}$ form a linearly independent set. (4.3.6)
  - ▷ The rows of $\mathbf{A}$ form a linearly independent set. (4.3.7)

Proof. By definition, the columns of $\mathbf{A}$ are a linearly independent set when the only set of α's satisfying the homogeneous equation

$$\mathbf{0} = \alpha_1\mathbf{A}_{*1} + \alpha_2\mathbf{A}_{*2} + \cdots + \alpha_n\mathbf{A}_{*n} = \left(\mathbf{A}_{*1} \,|\, \mathbf{A}_{*2} \,|\, \cdots \,|\, \mathbf{A}_{*n}\right)\begin{pmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{pmatrix}$$

is the trivial solution $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$, which is equivalent to saying $N(\mathbf{A}) = \{\mathbf{0}\}$. The fact that $N(\mathbf{A}) = \{\mathbf{0}\}$ is equivalent to $rank(\mathbf{A}) = n$ was demonstrated in (4.2.10). Statements (4.3.4) and (4.3.5) follow by replacing $\mathbf{A}$ by $\mathbf{A}^T$ in (4.3.2) and (4.3.3) and by using the fact that $rank(\mathbf{A}) = rank\left(\mathbf{A}^T\right)$. Statements (4.3.6) and (4.3.7) are simply special cases of (4.3.3) and (4.3.5). ∎

Example 4.3.2

Any set $\{\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_n}\}$ consisting of distinct unit vectors is a linearly independent set because $rank\left(\mathbf{e}_{i_1} \,|\, \mathbf{e}_{i_2} \,|\, \cdots \,|\, \mathbf{e}_{i_n}\right) = n$. For example, the set of unit vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_4\}$ in $\Re^4$ is linearly independent because $rank\begin{pmatrix}1&0&0\\0&1&0\\0&0&0\\0&0&1\end{pmatrix} = 3$.

Example 4.3.3

Diagonal Dominance. A matrix $\mathbf{A}_{n\times n}$ is said to be ***diagonally dominant*** whenever

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \quad \text{for each } i = 1, 2, \ldots, n.$$

That is, the magnitude of each diagonal entry exceeds the sum of the magnitudes of the off-diagonal entries in the corresponding row. Diagonally dominant matrices occur naturally in a wide variety of practical applications, and when solving a diagonally dominant system by Gaussian elimination, partial pivoting is never required—you are asked to provide the details in Exercise 4.3.15.

Problem: In 1900, Minkowski (p. 278) discovered that all diagonally dominant matrices are nonsingular. Establish the validity of Minkowski's result.

Solution: The strategy is to prove that if $\mathbf{A}$ is diagonally dominant, then $N(\mathbf{A}) = \{\mathbf{0}\}$, so that (4.3.2) together with (4.3.6) will provide the desired conclusion. Use an indirect argument—suppose there exists a vector $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{Ax} = \mathbf{0}$, and assume that x_k is the entry of maximum magnitude in $\mathbf{x}$. Focus on the kth component of $\mathbf{Ax}$, and write the equation $\mathbf{A}_{k*}\mathbf{x} = 0$ as

$$a_{kk}x_k = -\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j.$$

Taking absolute values of both sides and using the triangle inequality together with the fact that $|x_j| \leq |x_k|$ for each j produces

$$|a_{kk}|\,|x_k| = \left|\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j\right| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}x_j| = \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|\,|x_j| \leq \left(\sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|\right)|x_k|.$$

But this implies that

$$|a_{kk}| \leq \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|,$$

which violates the hypothesis that $\mathbf{A}$ is diagonally dominant. Therefore, the assumption that there exists a nonzero vector in $N(\mathbf{A})$ must be false, so we may conclude that $N(\mathbf{A}) = \{\mathbf{0}\}$, and hence $\mathbf{A}$ is nonsingular.

Note: An alternate solution is given in Example 7.1.6 on p. 499.

Example 4.3.4

Vandermonde Matrices. Matrices of the form

$$\mathbf{V}_{m\times n} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{n-1} \end{pmatrix}$$

in which $x_i \neq x_j$ for all $i \neq j$ are called ***Vandermonde***[26] ***matrices.***

Problem: Explain why the columns in $\mathbf{V}$ constitute a linearly independent set whenever $n \leq m$.

Solution: According to (4.3.2), the columns of $\mathbf{V}$ form a linearly independent set if and only if $N(\mathbf{V}) = \{\mathbf{0}\}$. If

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{n-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{4.3.8}$$

then for each $i = 1, 2, \ldots, m$,

$$\alpha_0 + x_i\alpha_1 + x_i^2\alpha_2 + \cdots + x_i^{n-1}\alpha_{n-1} = 0.$$

This implies that the polynomial

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{n-1}x^{n-1}$$

has m distinct roots—namely, the x_i's. However, $\deg p(x) \leq n-1$ and the fundamental theorem of algebra guarantees that if $p(x)$ is not the zero polynomial, then $p(x)$ can have at most $n-1$ distinct roots. Therefore, (4.3.8) holds if and only if $\alpha_i = 0$ for all i, and thus (4.3.2) insures that the columns of $\mathbf{V}$ form a linearly independent set.

[26] This is named in honor of the French mathematician Alexandre-Theophile Vandermonde (1735–1796). He made a variety of contributions to mathematics, but he is best known perhaps for being the first European to give a logically complete exposition of the theory of determinants. He is regarded by many as being the founder of that theory. However, the matrix $\mathbf{V}$ (and an associated determinant) named after him, by Lebesgue, does not appear in Vandermonde's published work. Vandermonde's first love was music, and he took up mathematics only after he was 35 years old. He advocated the theory that all art and music rested upon a general principle that could be expressed mathematically, and he claimed that almost anyone could become a composer with the aid of mathematics.

Example 4.3.5

Problem: Given a set of m points $\mathcal{S} = \{(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_m, y_m)\}$ in which the x_i's are distinct, explain why there is a unique polynomial

$$\ell(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_{m-1} t^{m-1} \tag{4.3.9}$$

of degree $m-1$ that passes through each point in $\mathcal{S}$.

Solution: The coefficients α_i must satisfy the equations

$$\begin{aligned}
&\alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 + \cdots + \alpha_{m-1} x_1^{m-1} = \ell(x_1) = y_1,\\
&\alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \cdots + \alpha_{m-1} x_2^{m-1} = \ell(x_2) = y_2,\\
&\vdots\\
&\alpha_0 + \alpha_1 x_m + \alpha_2 x_m^2 + \cdots + \alpha_{m-1} x_m^{m-1} = \ell(x_m) = y_m.
\end{aligned}$$

Writing this $m \times m$ system as

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{m-1} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

reveals that the coefficient matrix is a square Vandermonde matrix, so the result of Example 4.3.4 guarantees that it is nonsingular. Consequently, the system has a unique solution, and thus there is one and only one possible set of coefficients for the polynomial $\ell(t)$ in (4.3.9). In fact, $\ell(t)$ must be given by

$$\ell(t) = \sum_{i=1}^{m} \left(y_i \frac{\prod_{j \neq i}^{m} (t - x_j)}{\prod_{j \neq i}^{m} (x_i - x_j)} \right).$$

Verify this by showing that the right-hand side is indeed a polynomial of degree $m-1$ that passes through the points in $\mathcal{S}$. The polynomial $\ell(t)$ is known as the ***Lagrange***[27] ***interpolation polynomial*** of degree $m-1$.

If $rank\,(\mathbf{A}_{m\times n}) < n$, then the columns of $\mathbf{A}$ must be a dependent set—recall (4.3.3). For such matrices we often wish to extract a ***maximal linearly independent subset*** of columns—i.e., a linearly independent set containing as many columns from $\mathbf{A}$ as possible. Although there can be several ways to make such a selection, the basic columns in $\mathbf{A}$ always constitute one solution.

[27] Joseph Louis Lagrange (1736–1813), born in Turin, Italy, is considered by many to be one of the two greatest mathematicians of the eighteenth century—Euler is the other. Lagrange occupied Euler's vacated position in 1766 in Berlin at the court of Frederick the Great who wrote that "the greatest king in Europe" wishes to have at his court "the greatest mathematician of Europe." After 20 years, Lagrange left Berlin and eventually moved to France. Lagrange's mathematical contributions are extremely wide and deep, but he had a particularly strong influence on the way mathematical research evolved. He was the first of the top-class mathematicians to recognize the weaknesses in the foundations of calculus, and he was among the first to attempt a rigorous development.

Maximal Independent Subsets

If $rank\,(\mathbf{A}_{m\times n}) = r$, then the following statements hold.

- Any maximal independent subset of columns from $\mathbf{A}$ contains exactly r columns. (4.3.10)
- Any maximal independent subset of rows from $\mathbf{A}$ contains exactly r rows. (4.3.11)
- In particular, the r basic columns in $\mathbf{A}$ constitute one maximal independent subset of columns from $\mathbf{A}$. (4.3.12)

Proof. Exactly the same linear relationships that exist among the columns of $\mathbf{A}$ must also hold among the columns of $\mathbf{E_A}$ —by (3.9.6). This guarantees that a subset of columns from $\mathbf{A}$ is linearly independent if and only if the columns in the corresponding positions in $\mathbf{E_A}$ are an independent set. Let

$$\mathbf{C} = \big(\mathbf{c}_1 \,|\, \mathbf{c}_2 \,|\, \cdots \,|\, \mathbf{c}_k\big)$$

be a matrix that contains an independent subset of columns from $\mathbf{E_A}$ so that $rank\,(\mathbf{C}) = k$ —recall (4.3.3). Since each column in $\mathbf{E_A}$ is a combination of the r basic (unit) columns in $\mathbf{E_A}$, there are scalars β_{ij} such that $\mathbf{c}_j = \sum_{i=1}^{r} \beta_{ij}\mathbf{e}_i$ for $j = 1, 2, \ldots, k$. These equations can be written as the single matrix equation

$$\big(\mathbf{c}_1 \,|\, \mathbf{c}_2 \,|\, \cdots \,|\, \mathbf{c}_k\big) = \big(\mathbf{e}_1 \,|\, \mathbf{e}_2 \,|\, \cdots \,|\, \mathbf{e}_r\big) \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rk} \end{pmatrix}$$

or

$$\mathbf{C}_{m\times k} = \begin{pmatrix} \mathbf{I}_r \\ \mathbf{0} \end{pmatrix} \mathbf{B}_{r\times k} = \begin{pmatrix} \mathbf{B}_{r\times k} \\ \mathbf{0} \end{pmatrix}, \qquad \text{where} \qquad \mathbf{B} = [\beta_{ij}].$$

Consequently, $r \geq rank\,(\mathbf{C}) = k$, and therefore any independent subset of columns from $\mathbf{E_A}$ —and hence any independent set of columns from $\mathbf{A}$ —cannot contain more than r vectors. Because the r basic (unit) columns in $\mathbf{E_A}$ form an independent set, the r basic columns in $\mathbf{A}$ constitute an independent set. This proves (4.3.10) and (4.3.12). The proof of (4.3.11) follows from the fact that $rank\,(\mathbf{A}) = rank\,(\mathbf{A}^T)$ —recall (3.9.11). ∎

Basic Facts of Independence

For a nonempty set of vectors $\mathcal{S} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ in a space $\mathcal{V}$, the following statements are true.

- If $\mathcal{S}$ contains a linearly dependent subset, then $\mathcal{S}$ itself must be linearly dependent. (4.3.13)
- If $\mathcal{S}$ is linearly independent, then every subset of $\mathcal{S}$ is also linearly independent. (4.3.14)
- If $\mathcal{S}$ is linearly independent and if $\mathbf{v} \in \mathcal{V}$, then the ***extension set*** $\mathcal{S}_{ext} = \mathcal{S} \cup \{\mathbf{v}\}$ is linearly independent if and only if $\mathbf{v} \notin span\,(\mathcal{S})$. (4.3.15)
- If $\mathcal{S} \subseteq \Re^m$ and if $n > m$, then $\mathcal{S}$ must be linearly dependent. (4.3.16)

Proof of (4.3.13). Suppose that $\mathcal{S}$ contains a linearly dependent subset, and, for the sake of convenience, suppose that the vectors in $\mathcal{S}$ have been permuted so that this dependent subset is $\mathcal{S}_{dep} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. According to the definition of dependence, there must be scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$, not all of which are zero, such that $\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k = \mathbf{0}$. This means that we can write

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k + 0\mathbf{u}_{k+1} + \cdots + 0\mathbf{u}_n = \mathbf{0},$$

where not all of the scalars are zero, and hence $\mathcal{S}$ is linearly dependent.

Proof of (4.3.14). This is an immediate consequence of (4.3.13).

Proof of (4.3.15). If $\mathcal{S}_{ext}$ is linearly independent, then $\mathbf{v} \notin span\,(\mathcal{S})$, for otherwise $\mathbf{v}$ would be a combination of vectors from $\mathcal{S}$ thus forcing $\mathcal{S}_{ext}$ to be a dependent set. Conversely, suppose $\mathbf{v} \notin span\,(\mathcal{S})$. To prove that $\mathcal{S}_{ext}$ is linearly independent, consider a linear combination

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n + \alpha_{n+1}\mathbf{v} = \mathbf{0}. \tag{4.3.17}$$

It must be the case that $\alpha_{n+1} = 0$, for otherwise $\mathbf{v}$ would be a combination of vectors from $\mathcal{S}$. Consequently,

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{0}.$$

But this implies that

$$\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$$

because $\mathcal{S}$ is linearly independent. Therefore, the only solution for the α's in (4.3.17) is the trivial set, and hence $\mathcal{S}_{ext}$ must be linearly independent.

Proof of (4.3.16). This follows from (4.3.3) because if the $\mathbf{u}_i$'s are placed as columns in a matrix $\mathbf{A}_{m\times n}$, then $rank\,(\mathbf{A}) \leq m < n$. ∎

Example 4.3.6

Let $\mathcal{V}$ be the vector space of real-valued functions of a real variable, and let $\mathcal{S} = \{f_1(x), f_2(x), \ldots, f_n(x)\}$ be a set of functions that are $n-1$ times differentiable. The ***Wronski***[28] ***matrix*** is defined to be

$$\mathbf{W}(x) = \begin{pmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{pmatrix}.$$

Problem: If there is at least one point $x = x_0$ such that $\mathbf{W}(x_0)$ is nonsingular, prove that $\mathcal{S}$ must be a linearly independent set.

Solution: Suppose that

$$0 = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \cdots + \alpha_n f_n(x) \tag{4.3.18}$$

for all values of x. When $x = x_0$, it follows that

$$\begin{aligned} 0 &= \alpha_1 f_1(x_0) + \alpha_2 f_2(x_0) + \cdots + \alpha_n f_n(x_0), \\ 0 &= \alpha_1 f_1'(x_0) + \alpha_2 f_2'(x_0) + \cdots + \alpha_n f_n'(x_0), \\ &\vdots \\ 0 &= \alpha_1 f_1^{(n-1)}(x_0) + \alpha_2 f_2^{(n-1)}(x_0) + \cdots + \alpha_n f_n^{(n-1)}(x_0), \end{aligned}$$

which means that $\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} \in N\big(\mathbf{W}(x_0)\big)$. But $N\big(\mathbf{W}(x_0)\big) = \{\mathbf{0}\}$ because $\mathbf{W}(x_0)$ is nonsingular, and hence $\mathbf{v} = \mathbf{0}$. Therefore, the only solution for the α's in (4.3.18) is the trivial solution $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ thereby insuring that $\mathcal{S}$ is linearly independent.

---

[28] This matrix is named in honor of the Polish mathematician Jozef Maria Höené Wronski (1778–1853), who studied four special forms of determinants, one of which was the determinant of the matrix that bears his name. Wronski was born to a poor family near Poznan, Poland, but he studied in Germany and spent most of his life in France. He is reported to have been an egotistical person who wrote in an exhaustively wearisome style. Consequently, almost no one read his work. Had it not been for his lone follower, Ferdinand Schweins (1780–1856) of Heidelberg, Wronski would probably be unknown today. Schweins preserved and extended Wronski's results in his own writings, which in turn received attention from others. Wronski also wrote on philosophy. While trying to reconcile Kant's metaphysics with Leibniz's calculus, Wronski developed a social philosophy called "Messianism" that was based on the belief that absolute truth could be achieved through mathematics.

For example, to verify that the set of polynomials $\mathcal{P} = \{1, x, x^2, \ldots, x^n\}$ is linearly independent, observe that the associated Wronski matrix

$$\mathbf{W}(x) = \begin{pmatrix} 1 & x & x^2 & \cdots & x^n \\ 0 & 1 & 2x & \cdots & nx^{n-1} \\ 0 & 0 & 2 & \cdots & n(n-1)x^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n! \end{pmatrix}$$

is triangular with nonzero diagonal entries. Consequently, $\mathbf{W}(x)$ is nonsingular for every value of x, and hence $\mathcal{P}$ must be an independent set.

Exercises for section 4.3

4.3.1. Determine which of the following sets are linearly independent. For those sets that are linearly dependent, write one of the vectors as a linear combination of the others.

(a) $\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix} \right\}$,

(b) $\{(1 \quad 2 \quad 3), (0 \quad 4 \quad 5), (0 \quad 0 \quad 6), (1 \quad 1 \quad 1)\}$,

(c) $\left\{ \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix} \right\}$,

(d) $\{(2 \quad 2 \quad 2 \quad 2), (2 \quad 2 \quad 0 \quad 2), (2 \quad 0 \quad 2 \quad 2)\}$,

(e) $\left\{ \begin{pmatrix} 1 \\ 2 \\ 0 \\ 4 \\ 0 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \\ 4 \\ 1 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 1 \\ 4 \\ 0 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 0 \\ 4 \\ 0 \\ 3 \\ 1 \end{pmatrix} \right\}$.

4.3.2. Consider the matrix $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 2 & 1 & 2 \\ 6 & 3 & 2 & 2 \end{pmatrix}$.

(a) Determine a maximal linearly independent subset of columns from $\mathbf{A}$.

(b) Determine the total number of linearly independent subsets that can be constructed using the columns of $\mathbf{A}$.

4.3.3. Suppose that in a population of a million children the height of each one is measured at ages 1 year, 2 years, and 3 years, and accumulate this data in a matrix

$$\begin{array}{c} \\ \#1 \\ \#2 \\ \vdots \\ \#i \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccc} 1\text{ yr} & 2\text{ yr} & 3\text{ yr} \end{array} \\ \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ \vdots & \vdots & \vdots \\ h_{i1} & h_{i2} & h_{i3} \\ \vdots & \vdots & \vdots \end{pmatrix} \end{array} = \mathbf{H}.$$

Explain why there are at most three "independent children" in the sense that the heights of all the other children must be a combination of these "independent" ones.

4.3.4. Consider a particular species of wildflower in which each plant has several stems, leaves, and flowers, and for each plant let the following hold.

$S =$ the average stem length (in inches).
$L =$ the average leaf width (in inches).
$F =$ the number of flowers.

Four particular plants are examined, and the information is tabulated in the following matrix:

$$\mathbf{A} = \begin{array}{c} \\ \#1 \\ \#2 \\ \#3 \\ \#4 \end{array} \begin{array}{c} \begin{array}{ccc} S & L & F \end{array} \\ \begin{pmatrix} 1 & 1 & 10 \\ 2 & 1 & 12 \\ 2 & 2 & 15 \\ 3 & 2 & 17 \end{pmatrix} \end{array}.$$

For these four plants, determine whether or not there exists a linear relationship between S, L, and F. In other words, do there exist constants $\alpha_0, \alpha_1, \alpha_2,$ and α_3 such that $\alpha_0 + \alpha_1 S + \alpha_2 L + \alpha_3 F = 0$?

4.3.5. Let $\mathcal{S} = \{\mathbf{0}\}$ be the set containing only the zero vector.

(a) Explain why $\mathcal{S}$ must be linearly dependent.
(b) Explain why any set containing a zero vector must be linearly dependent.

4.3.6. If $\mathbf{T}$ is a triangular matrix in which each $t_{ii} \neq 0$, explain why the rows and columns of $\mathbf{T}$ must each be linearly independent sets.

4.3.7. Determine whether or not the following set of matrices is a linearly independent set:

$$\left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right\}.$$

4.3.8. Without doing any computation, determine whether the following matrix is singular or nonsingular:

$$\mathbf{A} = \begin{pmatrix} n & 1 & 1 & \cdots & 1 \\ 1 & n & 1 & \cdots & 1 \\ 1 & 1 & n & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n \end{pmatrix}_{n \times n}.$$

4.3.9. In theory, determining whether or not a given set is linearly independent is a well-defined problem with a straightforward solution. In practice, however, this problem is often not so well defined because it becomes clouded by the fact that we usually cannot use exact arithmetic, and contradictory conclusions may be produced depending upon the precision of the arithmetic. For example, let

$$\mathcal{S} = \left\{ \begin{pmatrix} .1 \\ .4 \\ .7 \end{pmatrix}, \begin{pmatrix} .2 \\ .5 \\ .8 \end{pmatrix}, \begin{pmatrix} .3 \\ .6 \\ .901 \end{pmatrix} \right\}.$$

(a) Use exact arithmetic to determine whether or not $\mathcal{S}$ is linearly independent.

(b) Use 3-digit arithmetic (without pivoting or scaling) to determine whether or not $\mathcal{S}$ is linearly independent.

4.3.10. If $\mathbf{A}_{m \times n}$ is a matrix such that $\sum_{j=1}^{n} a_{ij} = 0$ for each $i = 1, 2, \ldots, m$ (i.e., each row sum is 0), explain why the columns of $\mathbf{A}$ are a linearly dependent set, and hence $rank\,(\mathbf{A}) < n$.

4.3.11. If $\mathcal{S} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is a linearly independent subset of $\Re^{m \times 1}$, and if $\mathbf{P}_{m \times m}$ is a nonsingular matrix, explain why the set

$$\mathbf{P}(\mathcal{S}) = \{\mathbf{P}\mathbf{u}_1, \mathbf{P}\mathbf{u}_2, \ldots, \mathbf{P}\mathbf{u}_n\}$$

must also be a linearly independent set. Is this result still true if $\mathbf{P}$ is singular?

4.3.12. Suppose that $\mathcal{S} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is a set of vectors from $\Re^m$. Prove that $\mathcal{S}$ is linearly independent if and only if the set

$$\mathcal{S}' = \left\{ \mathbf{u}_1,\ \sum_{i=1}^{2} \mathbf{u}_i,\ \sum_{i=1}^{3} \mathbf{u}_i,\ \ldots,\ \sum_{i=1}^{n} \mathbf{u}_i \right\}$$

is linearly independent.

4.3.13. Which of the following sets of functions are linearly independent?

(a) $\{\sin x,\ \cos x,\ x \sin x\}$.

(b) $\{e^x,\ xe^x,\ x^2 e^x\}$.

(c) $\{\sin^2 x,\ \cos^2 x,\ \cos 2x\}$.

4.3.14. Prove that the converse of the statement given in Example 4.3.6 is false by showing that $\mathcal{S} = \{x^3,\ |x|^3\}$ is a linearly independent set, but the associated Wronski matrix $\mathbf{W}(x)$ is singular for all values of x.

4.3.15. If $\mathbf{A}^T$ is diagonally dominant, explain why partial pivoting is not needed when solving $\mathbf{Ax} = \mathbf{b}$ by Gaussian elimination. **Hint:** If after one step of Gaussian elimination we have

$$\mathbf{A} = \begin{pmatrix} \alpha & \mathbf{d}^T \\ \mathbf{c} & \mathbf{B} \end{pmatrix} \xrightarrow{\textit{one step}} \begin{pmatrix} \alpha & \mathbf{d}^T \\ \mathbf{0} & \mathbf{B} - \dfrac{\mathbf{cd}^T}{\alpha} \end{pmatrix},$$

show that $\mathbf{A}^T$ being diagonally dominant implies $\mathbf{X} = \left(\mathbf{B} - \dfrac{\mathbf{cd}^T}{\alpha}\right)^T$ must also be diagonally dominant.

4.4 BASIS AND DIMENSION

Recall from §4.1 that $\mathcal{S}$ is a spanning set for a space $\mathcal{V}$ if and only if every vector in $\mathcal{V}$ is a linear combination of vectors in $\mathcal{S}$. However, spanning sets can contain redundant vectors. For example, a subspace $\mathcal{L}$ defined by a line through the origin in $\Re^2$ may be spanned by any number of nonzero vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ in $\mathcal{L}$, but any one of the vectors $\{\mathbf{v}_i\}$ by itself will suffice. Similarly, a plane $\mathcal{P}$ through the origin in $\Re^3$ can be spanned in many different ways, but the parallelogram law indicates that a minimal spanning set need only be an independent set of two vectors from $\mathcal{P}$. These considerations motivate the following definition.

Basis

A linearly independent spanning set for a vector space $\mathcal{V}$ is called a ***basis*** for $\mathcal{V}$.

It can be proven that every vector space $\mathcal{V}$ possesses a basis—details for the case when $\mathcal{V} \subseteq \Re^m$ are asked for in the exercises. Just as in the case of spanning sets, a space can possess many different bases.

Example 4.4.1

- The unit vectors $\mathcal{S} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ in $\Re^n$ are a basis for $\Re^n$. This is called the ***standard basis*** for $\Re^n$.

- If $\mathbf{A}$ is an $n \times n$ nonsingular matrix, then the set of rows in $\mathbf{A}$ as well as the set of columns from $\mathbf{A}$ constitute a basis for $\Re^n$. For example, (4.3.3) insures that the columns of $\mathbf{A}$ are linearly independent, and we know they span $\Re^n$ because $R(\mathbf{A}) = \Re^n$ —recall Exercise 4.2.5(b).

- For the trivial vector space $\mathcal{Z} = \{\mathbf{0}\}$, there is no nonempty linearly independent spanning set. Consequently, the empty set is considered to be a basis for $\mathcal{Z}$.

- The set $\{1, x, x^2, \ldots, x^n\}$ is a basis for the vector space of polynomials having degree n or less.

- The infinite set $\{1, x, x^2, \ldots\}$ is a basis for the vector space of all polynomials. It should be clear that no finite basis is possible.

Spaces that possess a basis containing an infinite number of vectors are referred to as ***infinite-dimensional spaces,*** and those that have a finite basis are called ***finite-dimensional spaces.*** This is often a line of demarcation in the study of vector spaces. A complete theoretical treatment would include the analysis of infinite-dimensional spaces, but this text is primarily concerned with finite-dimensional spaces over the real or complex numbers. It can be shown that, in effect, this amounts to analyzing $\Re^n$ or $\mathcal{C}^n$ and their subspaces.

The original concern of this section was to try to eliminate redundancies from spanning sets so as to provide spanning sets containing a minimal number of vectors. The following theorem shows that a basis is indeed such a set.

Characterizations of a Basis

Let $\mathcal{V}$ be a subspace of $\Re^m$, and let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\} \subseteq \mathcal{V}$. The following statements are equivalent.

- $\mathcal{B}$ is a basis for $\mathcal{V}$. (4.4.1)
- $\mathcal{B}$ is a minimal spanning set for $\mathcal{V}$. (4.4.2)
- $\mathcal{B}$ is a maximal linearly independent subset of $\mathcal{V}$. (4.4.3)

Proof. First argue that (4.4.1) $\Longrightarrow$ (4.4.2) $\Longrightarrow$ (4.4.1), and then show (4.4.1) is equivalent to (4.4.3).

Proof of (4.4.1) $\Longrightarrow$ (4.4.2). First suppose that $\mathcal{B}$ is a basis for $\mathcal{V}$, and prove that $\mathcal{B}$ is a minimal spanning set by using an indirect argument—i.e., assume that $\mathcal{B}$ is *not* minimal, and show that this leads to a contradiction. If $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a basis for $\mathcal{V}$ in which $k < n$, then each $\mathbf{b}_j$ can be written as a combination of the $\mathbf{x}_i$'s. That is, there are scalars α_{ij} such that

$$\mathbf{b}_j = \sum_{i=1}^{k} \alpha_{ij}\mathbf{x}_i \qquad \text{for } j = 1, 2, \ldots, n. \tag{4.4.4}$$

If the $\mathbf{b}$'s and $\mathbf{x}$'s are placed as columns in matrices

$$\mathbf{B}_{m\times n} = \big(\mathbf{b}_1 \,|\, \mathbf{b}_2 \,|\, \cdots \,|\, \mathbf{b}_n\big) \quad \text{and} \quad \mathbf{X}_{m\times k} = \big(\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_k\big),$$

then (4.4.4) can be expressed as the matrix equation

$$\mathbf{B} = \mathbf{X}\mathbf{A}, \qquad \text{where,} \qquad \mathbf{A}_{k\times n} = [\alpha_{ij}]\,.$$

Since the rank of a matrix cannot exceed either of its size dimensions, and since $k < n$, we have that $rank\,(\mathbf{A}) \le k < n$, so that $N\,(\mathbf{A}) \neq \{\mathbf{0}\}$—recall (4.2.10). If $\mathbf{z} \neq \mathbf{0}$ is such that $\mathbf{A}\mathbf{z} = \mathbf{0}$, then $\mathbf{B}\mathbf{z} = \mathbf{0}$. But this is impossible because

the columns of $\mathbf{B}$ are linearly independent, and hence $N(\mathbf{B}) = \{\mathbf{0}\}$—recall (4.3.2). Therefore, the supposition that there exists a basis for $\mathcal{V}$ containing fewer than n vectors must be false, and we may conclude that $\mathcal{B}$ is indeed a minimal spanning set.

Proof of (4.4.2) $\Longrightarrow$ (4.4.1). If $\mathcal{B}$ is a minimal spanning set, then $\mathcal{B}$ must be a *linearly independent* spanning set. Otherwise, some $\mathbf{b}_i$ would be a linear combination of the other $\mathbf{b}$'s, and the set

$$\mathcal{B}' = \{\mathbf{b}_1, \ldots, \mathbf{b}_{i-1}, \mathbf{b}_{i+1}, \ldots, \mathbf{b}_n\}$$

would still span $\mathcal{V}$, but $\mathcal{B}'$ would contain fewer vectors than $\mathcal{B}$, which is impossible because $\mathcal{B}$ is a *minimal* spanning set.

Proof of (4.4.3) $\Longrightarrow$ (4.4.1). If $\mathcal{B}$ is a maximal linearly independent subset of $\mathcal{V}$, but not a basis for $\mathcal{V}$, then there exists a vector $\mathbf{v} \in \mathcal{V}$ such that $\mathbf{v} \notin span(\mathcal{B})$. This means that the extension set

$$\mathcal{B} \cup \{\mathbf{v}\} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n, \mathbf{v}\}$$

is linearly independent—recall (4.3.15). But this is impossible because $\mathcal{B}$ is a *maximal* linearly independent subset of $\mathcal{V}$. Therefore, $\mathcal{B}$ is a basis for $\mathcal{V}$.

Proof of (4.4.1) $\Longrightarrow$ (4.4.3). Suppose that $\mathcal{B}$ is a basis for $\mathcal{V}$, but not a maximal linearly independent subset of $\mathcal{V}$, and let

$$\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k\} \subseteq \mathcal{V}, \qquad \text{where} \qquad k > n$$

be a maximal linearly independent subset—recall that (4.3.16) insures the existence of such a set. The previous argument shows that $\mathcal{Y}$ must be a basis for $\mathcal{V}$. But this is impossible because we already know that a basis must be a minimal spanning set, and $\mathcal{B}$ is a spanning set containing fewer vectors than $\mathcal{Y}$. Therefore, $\mathcal{B}$ must be a maximal linearly independent subset of $\mathcal{V}$. ∎

Although a space $\mathcal{V}$ can have many different bases, the preceding result guarantees that all bases for $\mathcal{V}$ contain the same number of vectors. If $\mathcal{B}_1$ and $\mathcal{B}_2$ are each a basis for $\mathcal{V}$, then each is a minimal spanning set, and thus they must contain the same number of vectors. As we are about to see, this number is quite important.

Dimension

The ***dimension*** of a vector space $\mathcal{V}$ is defined to be

$$\begin{aligned} \dim \mathcal{V} &= \text{number of vectors in any basis for } \mathcal{V} \\ &= \text{number of vectors in any minimal spanning set for } \mathcal{V} \\ &= \text{number of vectors in any maximal independent subset of } \mathcal{V}. \end{aligned}$$

Example 4.4.2

- If $\mathcal{Z} = \{\mathbf{0}\}$ is the trivial subspace, then $\dim \mathcal{Z} = 0$ because the basis for this space is the empty set.
- If $\mathcal{L}$ is a line through the origin in $\Re^3$, then $\dim \mathcal{L} = 1$ because a basis for $\mathcal{L}$ consists of any nonzero vector lying along $\mathcal{L}$.
- If $\mathcal{P}$ is a plane through the origin in $\Re^3$, then $\dim \mathcal{P} = 2$ because a minimal spanning set for $\mathcal{P}$ must contain two vectors from $\mathcal{P}$.
- $\dim \Re^3 = 3$ because the three unit vectors $\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$ constitute a basis for $\Re^3$.
- $\dim \Re^n = n$ because the unit vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ in $\Re^n$ form a basis.

Example 4.4.3

Problem: If $\mathcal{V}$ is an n-dimensional space, explain why every independent subset $\mathcal{S} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subset \mathcal{V}$ containing n vectors must be a basis for $\mathcal{V}$.

Solution: $\dim \mathcal{V} = n$ means that every subset of $\mathcal{V}$ that contains more than n vectors must be linearly dependent. Consequently, $\mathcal{S}$ is a maximal independent subset of $\mathcal{V}$, and hence $\mathcal{S}$ is a basis for $\mathcal{V}$.

Example 4.4.2 shows that in a loose sense the dimension of a space is a measure of the amount of "stuff" in the space—a plane $\mathcal{P}$ in $\Re^3$ has more "stuff" in it than a line $\mathcal{L}$, but $\mathcal{P}$ contains less "stuff" than the entire space $\Re^3$. Recall from the discussion in §4.1 that subspaces of $\Re^n$ are generalized versions of flat surfaces through the origin. The concept of dimension gives us a way to distinguish between these "flat" objects according to how much "stuff" they contain—much the same way we distinguish between lines and planes in $\Re^3$. Another way to think about dimension is in terms of "degrees of freedom." In the trivial space $\mathcal{Z}$, there are no degrees of freedom—you can move nowhere—whereas on a line there is one degree of freedom—length; in a plane there are two degrees of freedom—length and width; in $\Re^3$ there are three degrees of freedom—length, width, and height; etc.

It is important not to confuse the dimension of a vector space $\mathcal{V}$ with the number of components contained in the individual vectors from $\mathcal{V}$. For example, if $\mathcal{P}$ is a plane through the origin in $\Re^3$, then $\dim \mathcal{P} = 2$, but the individual vectors in $\mathcal{P}$ each have three components. Although the dimension of a space $\mathcal{V}$ and the number of components contained in the individual vectors from $\mathcal{V}$ need not be the same, they are nevertheless related. For example, if $\mathcal{V}$ is a subspace of $\Re^n$, then (4.3.16) insures that no linearly independent subset in $\mathcal{V}$ can contain more than n vectors and, consequently, $\dim \mathcal{V} \leq n$. This observation generalizes to produce the following theorem.

Subspace Dimension

For vector spaces $\mathcal{M}$ and $\mathcal{N}$ such that $\mathcal{M} \subseteq \mathcal{N}$, the following statements are true.

- $\dim \mathcal{M} \leq \dim \mathcal{N}$. (4.4.5)
- If $\dim \mathcal{M} = \dim \mathcal{N}$, then $\mathcal{M} = \mathcal{N}$. (4.4.6)

Proof. Let $\dim \mathcal{M} = m$ and $\dim \mathcal{N} = n$, and use an indirect argument to prove (4.4.5). If it were the case that $m > n$, then there would exist a linearly independent subset of $\mathcal{N}$ (namely, a basis for $\mathcal{M}$) containing more than n vectors. But this is impossible because $\dim \mathcal{N}$ is the size of a maximal independent subset of $\mathcal{N}$. Thus $m \leq n$. Now prove (4.4.6). If $m = n$ but $\mathcal{M} \neq \mathcal{N}$, then there exists a vector $\mathbf{x}$ such that $\mathbf{x} \in \mathcal{N}$ but $\mathbf{x} \notin \mathcal{M}$. If $\mathcal{B}$ is a basis for $\mathcal{M}$, then $\mathbf{x} \notin span\,(\mathcal{B})$, and the extension set $\mathcal{E} = \mathcal{B} \cup \{\mathbf{x}\}$ is a linearly independent subset of $\mathcal{N}$—recall (4.3.15). But $\mathcal{E}$ contains $m + 1 = n + 1$ vectors, which is impossible because $\dim \mathcal{N} = n$ is the size of a maximal independent subset of $\mathcal{N}$. Hence $\mathcal{M} = \mathcal{N}$. ∎

Let's now find bases and dimensions for the four fundamental subspaces of an $m \times n$ matrix $\mathbf{A}$ of rank r, and let's start with $R(\mathbf{A})$. The entire set of columns in $\mathbf{A}$ spans $R(\mathbf{A})$, but they won't form a basis when there are dependencies among some of the columns. However, the set of *basic* columns in $\mathbf{A}$ is also a spanning set—recall (4.2.8)—and the basic columns always constitute a linearly independent set because no basic column can be a combination of other basic columns (otherwise it wouldn't be basic). So, the set of basic columns is a basis for $R(\mathbf{A})$, and, since there are r of them, $\dim R(\mathbf{A}) = r = rank\,(\mathbf{A})$.

Similarly, the entire set of rows in $\mathbf{A}$ spans $R\left(\mathbf{A}^T\right)$, but the set of all rows is not a basis when dependencies exist. Recall from (4.2.7) that if $\mathbf{U} = \begin{pmatrix} \mathbf{C}_{r\times n} \\ \mathbf{0} \end{pmatrix}$ is any row echelon form that is row equivalent to $\mathbf{A}$, then the rows of $\mathbf{C}$ span $R\left(\mathbf{A}^T\right)$. Since $rank\,(\mathbf{C}) = r$, (4.3.5) insures that the rows of $\mathbf{C}$ are linearly independent. Consequently, the rows in $\mathbf{C}$ are a basis for $R\left(\mathbf{A}^T\right)$, and, since there are r of them, $\dim R\left(\mathbf{A}^T\right) = r = rank\,(\mathbf{A})$. Older texts referred to $\dim R\left(\mathbf{A}^T\right)$ as the *row rank* of $\mathbf{A}$, while $\dim R(\mathbf{A})$ was called the *column rank* of $\mathbf{A}$, and it was a major task to prove that the row rank always agrees with the column rank. Notice that this is a consequence of the discussion above where it was observed that $\dim R\left(\mathbf{A}^T\right) = r = \dim R(\mathbf{A})$.

Turning to the nullspaces, let's first examine $N\left(\mathbf{A}^T\right)$. We know from (4.2.12) that if $\mathbf{P}$ is a nonsingular matrix such that $\mathbf{PA} = \mathbf{U}$ is in row echelon form, then the last $m - r$ rows in $\mathbf{P}$ span $N\left(\mathbf{A}^T\right)$. Because the set of rows in a nonsingular matrix is a linearly independent set, and because any subset

of an independent set is again independent—see (4.3.7) and (4.3.14)—it follows that the last $m-r$ rows in $\mathbf{P}$ are linearly independent, and hence they constitute a basis for $N\left(\mathbf{A}^T\right)$. And this implies $\dim N\left(\mathbf{A}^T\right) = m-r$ (i.e., the number of rows in $\mathbf{A}$ minus the rank of $\mathbf{A}$). Replacing $\mathbf{A}$ by $\mathbf{A}^T$ shows that $\dim N\left(\mathbf{A}^{T^T}\right) = \dim N(\mathbf{A})$ is the number of rows in $\mathbf{A}^T$ minus $rank\left(\mathbf{A}^T\right)$. But $rank\left(\mathbf{A}^T\right) = rank(\mathbf{A}) = r$, so $\dim N(\mathbf{A}) = n-r$. We deduced $\dim N(\mathbf{A})$ without exhibiting a specific basis, but a basis for $N(\mathbf{A})$ is easy to describe. Recall that the set $\mathcal{H}$ containing the $\mathbf{h}_i$'s appearing in the general solution (4.2.9) of $\mathbf{Ax} = \mathbf{0}$ spans $N(\mathbf{A})$. Since there are exactly $n-r$ vectors in $\mathcal{H}$, and since $\dim N(\mathbf{A}) = n-r$, $\mathcal{H}$ is a minimal spanning set, so, by (4.4.2), $\mathcal{H}$ must be a basis for $N(\mathbf{A})$. Below is a summary of facts uncovered above.

Fundamental Subspaces—Dimension and Bases

For an $m \times n$ matrix of real numbers such that $rank(\mathbf{A}) = r$,

- $\dim R(\mathbf{A}) = r,$ (4.4.7)
- $\dim N(\mathbf{A}) = n-r,$ (4.4.8)
- $\dim R\left(\mathbf{A}^T\right) = r,$ (4.4.9)
- $\dim N\left(\mathbf{A}^T\right) = m-r.$ (4.4.10)

Let $\mathbf{P}$ be a nonsingular matrix such that $\mathbf{PA} = \mathbf{U}$ is in row echelon form, and let $\mathcal{H}$ be the set of $\mathbf{h}_i$'s appearing in the general solution (4.2.9) of $\mathbf{Ax} = \mathbf{0}$.

- The basic columns of $\mathbf{A}$ form a basis for $R(\mathbf{A})$. (4.4.11)
- The nonzero rows of $\mathbf{U}$ form a basis for $R\left(\mathbf{A}^T\right)$. (4.4.12)
- The set $\mathcal{H}$ is a basis for $N(\mathbf{A})$. (4.4.13)
- The last $m-r$ rows of $\mathbf{P}$ form a basis for $N\left(\mathbf{A}^T\right)$. (4.4.14)

For matrices with complex entries, the above statements remain valid provided that $\mathbf{A}^T$ is replaced with $\mathbf{A}^*$.

Statements (4.4.7) and (4.4.8) combine to produce the following theorem.

Rank Plus Nullity Theorem

- $\dim R(\mathbf{A}) + \dim N(\mathbf{A}) = n$ for all $m \times n$ matrices. (4.4.15)

In loose terms, this is a kind of conservation law—it says that as the amount of "stuff" in $R(\mathbf{A})$ increases, the amount of "stuff" in $N(\mathbf{A})$ must decrease, and vice versa. The phrase *rank plus nullity* is used because $\dim R(\mathbf{A})$ is the rank of $\mathbf{A}$, and $\dim N(\mathbf{A})$ was traditionally known as the ***nullity of*** $\mathbf{A}$.

Example 4.4.4

Problem: Determine the dimension as well as a basis for the space spanned by

$$\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix} \right\}.$$

Solution 1: Place the vectors as columns in a matrix $\mathbf{A}$, and reduce

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 5 \\ 2 & 0 & 6 \\ 1 & 2 & 7 \end{pmatrix} \longrightarrow \mathbf{E_A} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Since $span(\mathcal{S}) = R(\mathbf{A})$, we have

$$\dim\big(span(\mathcal{S})\big) = \dim R(\mathbf{A}) = rank(\mathbf{A}) = 2.$$

The basic columns $\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \right\}$ are a basis for $R(\mathbf{A}) = span(\mathcal{S})$. Other bases are also possible. Examining $\mathbf{E_A}$ reveals that any two vectors in $\mathcal{S}$ form an independent set, and therefore any pair of vectors from $\mathcal{S}$ constitutes a basis for $span(\mathcal{S})$.

Solution 2: Place the vectors from $\mathcal{S}$ as rows in a matrix $\mathbf{B}$, and reduce $\mathbf{B}$ to row echelon form:

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 2 \\ 5 & 6 & 7 \end{pmatrix} \longrightarrow \mathbf{U} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

This time we have $span(\mathcal{S}) = R\left(\mathbf{B}^T\right)$, so that

$$\dim\big(span(\mathcal{S})\big) = \dim R\left(\mathbf{B}^T\right) = rank(\mathbf{B}) = rank(\mathbf{U}) = 2,$$

and a basis for $span(\mathcal{S}) = R\left(\mathbf{B}^T\right)$ is given by the nonzero rows in $\mathbf{U}$.

Example 4.4.5

Problem: If $\mathcal{S}_r = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ is a linearly independent subset of an n-dimensional space $\mathcal{V}$, where $r < n$, explain why it must be possible to find extension vectors $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ from $\mathcal{V}$ such that

$$\mathcal{S}_n = \{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$$

is a basis for $\mathcal{V}$.

Solution 1: $r < n$ means that $span\,(\mathcal{S}_r) \neq \mathcal{V}$, and hence there exists a vector $\mathbf{v}_{r+1} \in \mathcal{V}$ such that $\mathbf{v}_{r+1} \notin span\,(\mathcal{S}_r)$. The extension set $\mathcal{S}_{r+1} = \mathcal{S}_r \cup \{\mathbf{v}_{r+1}\}$ is an independent subset of $\mathcal{V}$ containing $r+1$ vectors—recall (4.3.15). Repeating this process generates independent subsets $\mathcal{S}_{r+2}, \mathcal{S}_{r+3}, \ldots,$ and eventually leads to a maximal independent subset $\mathcal{S}_n \subset \mathcal{V}$ containing n vectors.

Solution 2: The first solution shows that it is theoretically possible to find extension vectors, but the argument given is not much help in actually computing them. It is easy to remedy this situation. Let $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be any basis for $\mathcal{V}$, and place the given $\mathbf{v}_i$'s along with the $\mathbf{b}_i$'s as columns in a matrix

$$\mathbf{A} = \big(\mathbf{v}_1 \,|\, \cdots \,|\, \mathbf{v}_r \,|\, \mathbf{b}_1 \,|\, \cdots \,|\, \mathbf{b}_n\big).$$

Clearly, $R\,(\mathbf{A}) = \mathcal{V}$ so that the set of basic columns from $\mathbf{A}$ is a basis for $\mathcal{V}$. Observe that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ are basic columns in $\mathbf{A}$ because no one of these is a combination of preceding ones. Therefore, the remaining $n-r$ basic columns must be a subset of $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$—say they are $\{\mathbf{b}_{j_1}, \mathbf{b}_{j_2}, \ldots, \mathbf{b}_{j_{n-r}}\}$. The complete set of basic columns from $\mathbf{A}$, and a basis for $\mathcal{V}$, is the set

$$\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_r, \mathbf{b}_{j_1}, \ldots, \mathbf{b}_{j_{n-r}}\}.$$

For example, to extend the independent set

$$\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ -2 \end{pmatrix} \right\}$$

to a basis for $\Re^4$, append the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4\}$ to the vectors in $\mathcal{S}$, and perform the reduction

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \\ 2 & -2 & 0 & 0 & 0 & 1 \end{pmatrix} \longrightarrow \mathbf{E}_{\mathbf{A}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1/2 \end{pmatrix}.$$

This reveals that $\{\mathbf{A}_{*1}, \mathbf{A}_{*2}, \mathbf{A}_{*4}, \mathbf{A}_{*5}\}$ are the basic columns in $\mathbf{A}$, and therefore

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ -2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

is a basis for $\Re^4$ that contains $\mathcal{S}$.

Example 4.4.6

Rank and Connectivity. A set of points (or *nodes*), $\{N_1, N_2, \ldots, N_m\}$, together with a set of paths (or *edges*), $\{E_1, E_2, \ldots, E_n\}$, between the nodes is called a ***graph***. A *connected graph* is one in which there is a sequence of edges linking any pair of nodes, and a *directed graph* is one in which each edge has been assigned a direction. For example, the graph in Figure 4.4.1 is both connected and directed.

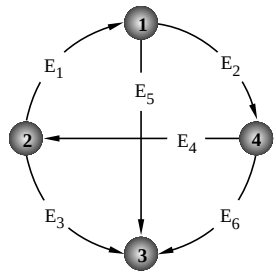


FIGURE 4.4.1

The connectivity of a directed graph is independent of the directions assigned to the edges—i.e., changing the direction of an edge doesn't change the connectivity. (Exercise 4.4.20 presents another type of connectivity in which direction matters.) On the surface, the concepts of graph connectivity and matrix rank seem to have little to do with each other, but, in fact, there is a close relationship. The ***incidence matrix*** associated with a directed graph containing m nodes and n edges is defined to be the $m \times n$ matrix $\mathbf{E}$ whose (k, j)-entry is

$$e_{kj} = \begin{cases} 1 & \text{if edge } E_j \text{ is directed } \textit{toward} \text{ node } N_k. \\ -1 & \text{if edge } E_j \text{ is directed } \textit{away from} \text{ node } N_k. \\ 0 & \text{if edge } E_j \text{ neither begins nor ends at node } N_k. \end{cases}$$

For example, the incidence matrix associated with the graph in Figure 4.4.1 is

$$\mathbf{E} = \begin{array}{c} \\ N_1 \\ N_2 \\ N_3 \\ N_4 \end{array} \begin{array}{c} \begin{array}{cccccc} E_1 & E_2 & E_3 & E_4 & E_5 & E_6 \end{array} \\ \begin{pmatrix} 1 & -1 & 0 & 0 & -1 & 0 \\ -1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & -1 & 0 & -1 \end{pmatrix} \end{array}. \tag{4.4.16}$$

Each edge in a directed graph is associated with two nodes—the nose and the tail of the edge—so each column in $\mathbf{E}$ must contain exactly two nonzero entries—a $(+1)$ and a (-1). Consequently, all column sums are zero. In other words, if $\mathbf{e}^T = (1 \quad 1 \quad \cdots \quad 1)$, then $\mathbf{e}^T\mathbf{E} = \mathbf{0}$, so $\mathbf{e} \in N\left(\mathbf{E}^T\right)$, and

$$rank\,(\mathbf{E}) = rank\left(\mathbf{E}^T\right) = m - \dim N\left(\mathbf{E}^T\right) \leq m - 1. \tag{4.4.17}$$

This inequality holds regardless of the connectivity of the associated graph, but marvelously, equality is attained if and only if the graph is connected.

Rank and Connectivity

Let $\mathcal{G}$ be a graph containing m nodes. If $\mathcal{G}$ is undirected, arbitrarily assign directions to the edges to make $\mathcal{G}$ directed, and let $\mathbf{E}$ be the corresponding incidence matrix.

- $\mathcal{G}$ is connected if and only if $rank\,(\mathbf{E}) = m - 1$. (4.4.18)

Proof. Suppose $\mathcal{G}$ is connected. Prove $rank\,(\mathbf{E}) = m - 1$ by arguing that $\dim N\left(\mathbf{E}^T\right) = 1$, and do so by showing $\mathbf{e} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$ is a basis $N\left(\mathbf{E}^T\right)$. To see that $\mathbf{e}$ spans $N\left(\mathbf{E}^T\right)$, consider an arbitrary $\mathbf{x} \in N\left(\mathbf{E}^T\right)$, and focus on any two components x_i and x_k in $\mathbf{x}$ along with the corresponding nodes N_i and N_k in $\mathcal{G}$. Since $\mathcal{G}$ is connected, there must exist a subset of r nodes,

$$\{N_{j_1}, N_{j_2}, \ldots, N_{j_r}\}, \quad \text{where} \quad i = j_1 \quad \text{and} \quad k = j_r,$$

such that there is an edge between N_{j_p} and $N_{j_{p+1}}$ for each $p = 1, 2, \ldots, r-1$. Therefore, corresponding to each of the $r-1$ pairs $\left(N_{j_p}, N_{j_{p+1}}\right)$, there must exist a column $\mathbf{c}_p$ in $\mathbf{E}$ (not necessarily the pth column) such that components j_p and j_{p+1} in $\mathbf{c}_p$ are complementary in the sense that one is $(+1)$ while the other is (-1) (all other components are zero). Because $\mathbf{x}^T\mathbf{E} = \mathbf{0}$, it follows that $\mathbf{x}^T\mathbf{c}_p = 0$, and hence $x_{j_p} = x_{j_{p+1}}$. But this holds for every $p = 1, 2, \ldots, r-1$, so $x_i = x_k$ for each i and k, and hence $\mathbf{x} = \alpha\mathbf{e}$ for some scalar α. Thus $\{\mathbf{e}\}$ spans $N\left(\mathbf{E}^T\right)$. Clearly, $\{\mathbf{e}\}$ is linearly independent, so it is a basis $N\left(\mathbf{E}^T\right)$, and, therefore, $\dim N\left(\mathbf{E}^T\right) = 1$ or, equivalently, $rank\,(\mathbf{E}) = m-1$. Conversely, suppose $rank\,(\mathbf{E}) = m-1$, and prove $\mathcal{G}$ is connected with an indirect argument. If $\mathcal{G}$ is not connected, then $\mathcal{G}$ is decomposable into two nonempty subgraphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in which there are no edges between nodes in $\mathcal{G}_1$ and nodes in $\mathcal{G}_2$. This means that the nodes in $\mathcal{G}$ can be ordered so as to make $\mathbf{E}$ have the form

$$\mathbf{E} = \begin{pmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_2 \end{pmatrix},$$

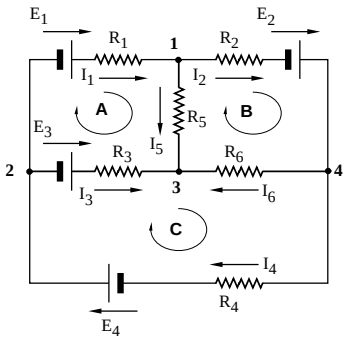
where $\mathbf{E}_1$ and $\mathbf{E}_2$ are the incidence matrices for $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. If $\mathcal{G}_1$ and $\mathcal{G}_2$ contain m_1 and m_2 nodes, respectively, then (4.4.17) insures that

$$rank\,(\mathbf{E}) = rank\begin{pmatrix} \mathbf{E}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_2 \end{pmatrix} = rank\,(\mathbf{E}_1) + rank\,(\mathbf{E}_1) \leq (m_1 - 1) + (m_2 - 1) = m - 2.$$

But this contradicts the hypothesis that $rank\,(\mathbf{E}) = m - 1$, so the supposition that $\mathcal{G}$ is not connected must be false. ∎

Example 4.4.7

An Application to Electrical Circuits. Recall from the discussion on p. 73 that applying Kirchhoff's node rule to an electrical circuit containing m nodes and n branches produces m homogeneous linear equations in n unknowns (the branch currents), and Kirchhoff's loop rule provides a nonhomogeneous equation for each simple loop in the circuit. For example, consider the circuit in Figure 4.4.2 along with its four nodal equations and three loop equations—this is the same circuit appearing on p. 73, and the equations are derived there.



$$\begin{aligned}
\text{Node 1:}\quad & I_1 - I_2 - I_5 = 0\\
\text{Node 2:}\quad & -I_1 - I_3 + I_4 = 0\\
\text{Node 3:}\quad & I_3 + I_5 + I_6 = 0\\
\text{Node 4:}\quad & I_2 - I_4 - I_6 = 0\\
\text{Loop A:}\quad & I_1R_1 - I_3R_3 + I_5R_5 = E_1 - E_3\\
\text{Loop B:}\quad & I_2R_2 - I_5R_5 + I_6R_6 = E_2\\
\text{Loop C:}\quad & I_3R_3 + I_4R_4 - I_6R_6 = E_3 + E_4
\end{aligned}$$

FIGURE 4.4.2

The directed graph and associated incidence matrix $\mathbf{E}$ defined by this circuit are the same as those appearing in Example 4.4.6 in Figure 4.4.1 and equation (4.4.16), so it's apparent that the 4×3 homogeneous system of nodal equations is precisely the system $\mathbf{Ex} = \mathbf{0}$. This observation holds for general circuits. The goal is to compute the six currents $I_1, I_2, \ldots, I_6$ by selecting six independent equations from the entire set of node and loop equations. In general, if a circuit containing m nodes is connected in the graph sense, then (4.4.18) insures that $rank\,(\mathbf{E}) = m - 1$, so there are m independent nodal equations. But Example 4.4.6 also shows that $\mathbf{0} = \mathbf{e}^T\mathbf{E} = \mathbf{E}_{1*} + \mathbf{E}_{2*} + \cdots + \mathbf{E}_{m*}$, which means that any row can be written in terms of the others, and this in turn implies that *every* subset of $m - 1$ rows in $\mathbf{E}$ must be independent (see Exercise 4.4.13). Consequently, when any nodal equation is discarded, the remaining ones are guaranteed to be independent. To determine an $n \times n$ nonsingular system that has the n branch currents as its unique solution, it's therefore necessary to find $n - m + 1$ additional independent equations, and, as shown in §2.6, these are the loop equations. A simple loop in a circuit is now seen to be a connected subgraph that does not properly contain other connected subgraphs. Physics dictates that the currents must be uniquely determined, so there must always be $n - m + 1$ simple loops, and the combination of these loop equations together with any subset of $m - 1$ nodal equations will be a nonsingular $n \times n$ system that yields the branch currents as its unique solution. For example, any three of the nodal equations in Figure 4.4.2 can be coupled with the three simple loop equations to produce a 6×6 nonsingular system whose solution is the six branch currents.

If $\mathcal{X}$ and $\mathcal{Y}$ are subspaces of a vector space $\mathcal{V}$, then the sum of $\mathcal{X}$ and $\mathcal{Y}$ was defined in §4.1 to be

$$\mathcal{X} + \mathcal{Y} = \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{X} \text{ and } \mathbf{y} \in \mathcal{Y}\},$$

and it was demonstrated in (4.1.1) that $\mathcal{X} + \mathcal{Y}$ is again a subspace of $\mathcal{V}$. You were asked in Exercise 4.1.8 to prove that the intersection $\mathcal{X} \cap \mathcal{Y}$ is also a subspace of $\mathcal{V}$. We are now in a position to exhibit an important relationship between $\dim(\mathcal{X} + \mathcal{Y})$ and $\dim(\mathcal{X} \cap \mathcal{Y})$.

Dimension of a Sum

If $\mathcal{X}$ and $\mathcal{Y}$ are subspaces of a vector space $\mathcal{V}$, then

$$\dim(\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim(\mathcal{X} \cap \mathcal{Y}). \tag{4.4.19}$$

Proof. The strategy is to construct a basis for $\mathcal{X} + \mathcal{Y}$ and count the number of vectors it contains. Let $\mathcal{S} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t\}$ be a basis for $\mathcal{X} \cap \mathcal{Y}$. Since $\mathcal{S} \subseteq \mathcal{X}$ and $\mathcal{S} \subseteq \mathcal{Y}$, there must exist extension vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ such that

$$\mathcal{B}_X = \{\mathbf{z}_1, \ldots, \mathbf{z}_t, \mathbf{x}_1, \ldots, \mathbf{x}_m\} = \text{ a basis for } \mathcal{X}$$

and

$$\mathcal{B}_Y = \{\mathbf{z}_1, \ldots, \mathbf{z}_t, \mathbf{y}_1, \ldots, \mathbf{y}_n\} = \text{ a basis for } \mathcal{Y}.$$

We know from (4.1.2) that $\mathcal{B} = \mathcal{B}_X \cup \mathcal{B}_Y$ spans $\mathcal{X} + \mathcal{Y}$, and we wish show that $\mathcal{B}$ is linearly independent. If

$$\sum_{i=1}^{t} \alpha_i \mathbf{z}_i + \sum_{j=1}^{m} \beta_j \mathbf{x}_j + \sum_{k=1}^{n} \gamma_k \mathbf{y}_k = \mathbf{0}, \tag{4.4.20}$$

then

$$\sum_{k=1}^{n} \gamma_k \mathbf{y}_k = -\left(\sum_{i=1}^{t} \alpha_i \mathbf{z}_i + \sum_{j=1}^{m} \beta_j \mathbf{x}_j \right) \in \mathcal{X}.$$

Since it is also true that $\sum_k \gamma_k \mathbf{y}_k \in \mathcal{Y}$, we have that $\sum_k \gamma_k \mathbf{y}_k \in \mathcal{X} \cap \mathcal{Y}$, and hence there must exist scalars δ_i such that

$$\sum_{k=1}^{n} \gamma_k \mathbf{y}_k = \sum_{i=1}^{t} \delta_i \mathbf{z}_i \quad \text{or, equivalently,} \quad \sum_{k=1}^{n} \gamma_k \mathbf{y}_k - \sum_{i=1}^{t} \delta_i \mathbf{z}_i = \mathbf{0}.$$

Since $\mathcal{B}_Y$ is an independent set, it follows that all of the γ_k's (as well as all δ_i's) are zero, and (4.4.20) reduces to $\sum_{i=1}^{t} \alpha_i \mathbf{z}_i + \sum_{j=1}^{m} \beta_j \mathbf{x}_j = \mathbf{0}$. But $\mathcal{B}_X$ is also an independent set, so the only way this can hold is for all of the α_i's as well as all of the β_j's to be zero. Therefore, the only possible solution for the α's, β's, and γ's in the homogeneous equation (4.4.20) is the trivial solution, and thus $\mathcal{B}$ is linearly independent. Since $\mathcal{B}$ is an independent spanning set, it is a basis for $\mathcal{X} + \mathcal{Y}$ and, consequently,

$$\dim(\mathcal{X} + \mathcal{Y}) = t+m+n = (t+m)+(t+n)-t = \dim \mathcal{X} + \dim \mathcal{Y} - \dim(\mathcal{X} \cap \mathcal{Y}). \quad \blacksquare$$

Example 4.4.8

Problem: Show that $rank\,(\mathbf{A}+\mathbf{B}) \leq rank\,(\mathbf{A}) + rank\,(\mathbf{B})$.

Solution: Observe that

$$R(\mathbf{A}+\mathbf{B}) \subseteq R(\mathbf{A}) + R(\mathbf{B})$$

because if $\mathbf{b} \in R(\mathbf{A}+\mathbf{B})$, then there is a vector $\mathbf{x}$ such that

$$\mathbf{b} = (\mathbf{A}+\mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x} \in R(\mathbf{A}) + R(\mathbf{B}).$$

Recall from (4.4.5) that if $\mathcal{M}$ and $\mathcal{N}$ are vector spaces such that $\mathcal{M} \subseteq \mathcal{N}$, then $\dim \mathcal{M} \leq \dim \mathcal{N}$. Use this together with formula (4.4.19) for the dimension of a sum to conclude that

$$\begin{aligned} rank\,(\mathbf{A}+\mathbf{B}) &= \dim R(\mathbf{A}+\mathbf{B}) \leq \dim\Big(R(\mathbf{A}) + R(\mathbf{B})\Big) \\ &= \dim R(\mathbf{A}) + \dim R(\mathbf{B}) - \dim\Big(R(\mathbf{A}) \cap R(\mathbf{B})\Big) \\ &\leq \dim R(\mathbf{A}) + \dim R(\mathbf{B}) = rank\,(\mathbf{A}) + rank\,(\mathbf{B}). \end{aligned}$$

Exercises for section 4.4

4.4.1. Find the dimensions of the four fundamental subspaces associated with

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

4.4.2. Find a basis for each of the four fundamental subspaces associated with

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

4.4.3. Determine the dimension of the space spanned by the set

$$\mathcal{S} = \left\{ \begin{pmatrix} 1 \\ 2 \\ -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 \\ 8 \\ -4 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 3 \\ 0 \\ 6 \end{pmatrix} \right\}.$$

4.4.4. Determine the dimensions of each of the following vector spaces:
(a) The space of polynomials having degree n or less.
(b) The space $\Re^{m \times n}$ of $m \times n$ matrices.
(c) The space of $n \times n$ symmetric matrices.

4.4.5. Consider the following matrix and column vector:

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

Verify that $\mathbf{v} \in N(\mathbf{A})$, and then extend $\{\mathbf{v}\}$ to a basis for $N(\mathbf{A})$.

4.4.6. Determine whether or not the set

$$\mathcal{B} = \left\{ \begin{pmatrix} 2 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\}$$

is a basis for the space spanned by the set

$$\mathcal{A} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 8 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \\ 1 \end{pmatrix} \right\}.$$

4.4.7. Construct a 4×4 homogeneous system of equations that has no zero coefficients and three linearly independent solutions.

4.4.8. Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be a basis for a vector space $\mathcal{V}$. Prove that each $\mathbf{v} \in \mathcal{V}$ can be expressed as a linear combination of the $\mathbf{b}_i$'s

$$\mathbf{v} = \alpha_1 \mathbf{b}_1 + \alpha_2 \mathbf{b}_2 + \cdots + \alpha_n \mathbf{b}_n,$$

in only one way—i.e., the *coordinates* α_i are unique.

4.4.9. For $\mathbf{A} \in \Re^{m\times n}$ and a subspace $\mathcal{S}$ of $\Re^{n\times 1}$, the image

$$\mathbf{A}(\mathcal{S}) = \{\mathbf{A}\mathbf{x} \,|\, \mathbf{x} \in \mathcal{S}\}$$

of $\mathcal{S}$ under $\mathbf{A}$ is a subspace of $\Re^{m\times 1}$—recall Exercise 4.1.9. Prove that if $\mathcal{S} \cap N(\mathbf{A}) = \mathbf{0}$, then $\dim \mathbf{A}(\mathcal{S}) = \dim(\mathcal{S})$. **Hint:** Use a basis $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ for $\mathcal{S}$ to determine a basis for $\mathbf{A}(\mathcal{S})$.

4.4.10. Explain why $\big|rank\,(\mathbf{A}) - rank\,(\mathbf{B})\big| \le rank\,(\mathbf{A} - \mathbf{B})$.

4.4.11. If $rank\,(\mathbf{A}_{m\times n}) = r$ and $rank\,(\mathbf{E}_{m\times n}) = k \le r$, explain why

$$r - k \le rank\,(\mathbf{A} + \mathbf{E}) \le r + k.$$

In words, this says that a perturbation of rank k can change the rank by at most k.

4.4.12. Explain why every nonzero subspace $\mathcal{V} \subseteq \Re^n$ must possess a basis.

4.4.13. Explain why *every* set of $m-1$ rows in the incidence matrix $\mathbf{E}$ of a connected directed graph containing m nodes is linearly independent.

4.4.14. For the incidence matrix $\mathbf{E}$ of a directed graph, explain why

$$\big[\mathbf{E}\mathbf{E}^T\big]_{ij} = \begin{cases} \text{number of edges at node } i & \text{when } i = j, \\ -(\text{number of edges between nodes } i \text{ and } j) & \text{when } i \neq j. \end{cases}$$

4.4.15. If $\mathcal{M}$ and $\mathcal{N}$ are subsets of a space $\mathcal{V}$, explain why

$$\dim\big(span\,(\mathcal{M} \cup \mathcal{N})\big) = \dim\big(span\,(\mathcal{M})\big) + \dim\big(span\,(\mathcal{N})\big) - \dim\big(span\,(\mathcal{M}) \cap span\,(\mathcal{N})\big).$$

4.4.16. Consider two matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{m\times k}$.

(a) Explain why

$$rank\,(\mathbf{A} \mid \mathbf{B}) = rank\,(\mathbf{A}) + rank\,(\mathbf{B}) - \dim\big(R(\mathbf{A}) \cap R(\mathbf{B})\big).$$

Hint: Recall Exercise 4.2.9.

(b) Now explain why

$$\dim N(\mathbf{A} \mid \mathbf{B}) = \dim N(\mathbf{A}) + \dim N(\mathbf{B}) + \dim\big(R(\mathbf{A}) \cap R(\mathbf{B})\big).$$

(c) Determine $\dim\big(R(\mathbf{C}) \cap N(\mathbf{C})\big)$ and $\dim\big(R(\mathbf{C}) + N(\mathbf{C})\big)$ for

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

4.4.17. Suppose that $\mathbf{A}$ is a matrix with m rows such that the system $\mathbf{Ax} = \mathbf{b}$ has a unique solution for every $\mathbf{b} \in \Re^m$. Explain why this means that $\mathbf{A}$ must be square and nonsingular.

4.4.18. Let $\mathcal{S}$ be the solution set for a consistent system of linear equations $\mathbf{Ax} = \mathbf{b}$.

(a) If $\mathcal{S}_{max} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_t\}$ is a maximal independent subset of $\mathcal{S}$, and if $\mathbf{p}$ is any particular solution, prove that

$$span\,(\mathcal{S}_{max}) = span\,\{\mathbf{p}\} + N\,(\mathbf{A}).$$

Hint: First show that $\mathbf{x} \in \mathcal{S}$ implies $\mathbf{x} \in span\,(\mathcal{S}_{max})$, and then demonstrate set inclusion in both directions with the aid of Exercise 4.2.10.

(b) If $\mathbf{b} \neq \mathbf{0}$ and $rank\,(\mathbf{A}_{m\times n}) = r$, explain why $\mathbf{Ax} = \mathbf{b}$ has $n - r + 1$ "independent solutions."

4.4.19. Let $rank\,(\mathbf{A}_{m\times n}) = r$, and suppose $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is a consistent system. If $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-r}\}$ is a basis for $N\,(\mathbf{A})$, and if $\mathbf{p}$ is a particular solution to $\mathbf{Ax} = \mathbf{b}$, show that

$$\mathcal{S}_{max} = \{\mathbf{p},\ \mathbf{p} + \mathbf{h}_1,\ \mathbf{p} + \mathbf{h}_2,\ \ldots,\ \mathbf{p} + \mathbf{h}_{n-r}\}$$

is a maximal independent set of solutions.

4.4.20. **Strongly Connected Graphs.** In Example 4.4.6 we started with a graph to construct a matrix, but it's also possible to reverse the situation by starting with a matrix to build an associated graph. The graph of $\mathbf{A}_{n\times n}$ (denoted by $\mathcal{G}(\mathbf{A})$) is defined to be the directed graph on n nodes $\{N_1, N_2, \ldots, N_n\}$ in which there is a directed edge leading *from* N_i *to* N_j if and only if $a_{ij} \neq 0$. The directed graph $\mathcal{G}(\mathbf{A})$ is said to be ***strongly connected*** provided that for each pair of nodes (N_i, N_k) there is a sequence of directed edges leading *from* N_i *to* N_k. The matrix $\mathbf{A}$ is said to be ***reducible*** if there exists a permutation matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$, where $\mathbf{X}$ and $\mathbf{Z}$ are both square matrices. Otherwise, $\mathbf{A}$ is said to be ***irreducible.*** Prove that $\mathcal{G}(\mathbf{A})$ is strongly connected if and only if $\mathbf{A}$ is irreducible. **Hint:** Prove the contrapositive: $\mathcal{G}(\mathbf{A})$ is *not* strongly connected if and only if $\mathbf{A}$ is reducible.

4.5 MORE ABOUT RANK

Since equivalent matrices have the same rank, it follows that if $\mathbf{P}$ and $\mathbf{Q}$ are nonsingular matrices such that the product $\mathbf{PAQ}$ is defined, then

$$rank\,(\mathbf{A}) = rank\,(\mathbf{PAQ}) = rank\,(\mathbf{PA}) = rank\,(\mathbf{AQ}).$$

In other words, rank is invariant under multiplication by a nonsingular matrix. However, multiplication by rectangular or singular matrices can alter the rank, and the following formula shows exactly how much alteration occurs.

Rank of a Product

If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then

$$rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}). \tag{4.5.1}$$

Proof. Start with a basis $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s\}$ for $N\,(\mathbf{A}) \cap R\,(\mathbf{B})$, and notice $N\,(\mathbf{A}) \cap R\,(\mathbf{B}) \subseteq R\,(\mathbf{B})$. If $\dim R\,(\mathbf{B}) = s + t$, then, as discussed in Example 4.4.5, there exists an extension set $\mathcal{S}_{ext} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t\}$ such that $\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_s, \mathbf{z}_1, \ldots, \mathbf{z}_t\}$ is a basis for $R\,(\mathbf{B})$. The goal is to prove that $\dim R\,(\mathbf{AB}) = t$, and this is done by showing $\mathcal{T} = \{\mathbf{Az}_1, \mathbf{Az}_2, \ldots, \mathbf{Az}_t\}$ is a basis for $R\,(\mathbf{AB})$. $\mathcal{T}$ spans $R\,(\mathbf{AB})$ because if $\mathbf{b} \in R\,(\mathbf{AB})$, then $\mathbf{b} = \mathbf{ABy}$ for some $\mathbf{y}$, but $\mathbf{By} \in R\,(\mathbf{B})$ implies $\mathbf{By} = \sum_{i=1}^{s} \xi_i \mathbf{x}_i + \sum_{i=1}^{t} \eta_i \mathbf{z}_i$, so

$$\mathbf{b} = \mathbf{A}\left(\sum_{i=1}^{s} \xi_i \mathbf{x}_i + \sum_{i=1}^{t} \eta_i \mathbf{z}_i\right) = \sum_{i=1}^{s} \xi_i \mathbf{Ax}_i + \sum_{i=1}^{t} \eta_i \mathbf{Az}_i = \sum_{i=1}^{t} \eta_i \mathbf{Az}_i.$$

$\mathcal{T}$ is linearly independent because if $\mathbf{0} = \sum_{i=1}^{t} \alpha_i \mathbf{Az}_i = \mathbf{A} \sum_{i=1}^{t} \alpha_i \mathbf{z}_i$, then $\sum_{i=1}^{t} \alpha_i \mathbf{z}_i \in N\,(\mathbf{A}) \cap R\,(\mathbf{B})$, so there are scalars β_j such that

$$\sum_{i=1}^{t} \alpha_i \mathbf{z}_i = \sum_{j=1}^{s} \beta_j \mathbf{x}_j \quad \text{or, equivalently,} \quad \sum_{i=1}^{t} \alpha_i \mathbf{z}_i - \sum_{j=1}^{s} \beta_j \mathbf{x}_j = \mathbf{0},$$

and hence the only solution for the α_i's and β_i's is the trivial solution because $\mathcal{B}$ is an independent set. Thus $\mathcal{T}$ is a basis for $R\,(\mathbf{AB})$, so $t = \dim R\,(\mathbf{AB}) = rank\,(\mathbf{AB})$, and hence

$$rank\,(\mathbf{B}) = \dim R\,(\mathbf{B}) = s + t = \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) + rank\,(\mathbf{AB}). \quad \blacksquare$$

It's sometimes necessary to determine an explicit basis for $N\,(\mathbf{A}) \cap R\,(\mathbf{B})$. In particular, such a basis is needed to construct the Jordan chains that are associated with the Jordan form that is discussed on pp. 582 and 594. The following example outlines a procedure for finding such a basis.

Basis for an Intersection

If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then a basis for $N(\mathbf{A}) \cap R(\mathbf{B})$ can be constructed by the following procedure.

- ▷ Find a basis $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ for $R(\mathbf{B})$.
- ▷ Set $\mathbf{X}_{n\times r} = \big(\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_r\big)$.
- ▷ Find a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ for $N(\mathbf{AX})$.
- ▷ $\mathcal{B} = \{\mathbf{Xv}_1, \mathbf{Xv}_2, \ldots, \mathbf{Xv}_s\}$ is a basis for $N(\mathbf{A}) \cap R(\mathbf{B})$.

Proof. The strategy is to argue that $\mathcal{B}$ is a maximal linear independent subset of $N(\mathbf{A}) \cap R(\mathbf{B})$. Since each $\mathbf{Xv}_j$ belongs to $R(\mathbf{X}) = R(\mathbf{B})$, and since $\mathbf{AXv}_j = \mathbf{0}$, it's clear that $\mathcal{B} \subset N(\mathbf{A}) \cap R(\mathbf{B})$. Let $\mathbf{V}_{r\times s} = \big(\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_s\big)$, and notice that $\mathbf{V}$ and $\mathbf{X}$ each have full column rank. Consequently, $N(\mathbf{X}) = \mathbf{0}$ so, by (4.5.1),

$$rank\,(\mathbf{XV})_{n\times s} = rank\,(\mathbf{V}) - \dim N(\mathbf{X}) \cap R(\mathbf{V}) = rank\,(\mathbf{V}) = s,$$

which insures that $\mathcal{B}$ is linearly independent. $\mathcal{B}$ is a *maximal* independent subset of $N(\mathbf{A}) \cap R(\mathbf{B})$ because (4.5.1) also guarantees that

$$\begin{aligned} s &= \dim N(\mathbf{AX}) = \dim N(\mathbf{X}) + \dim N(\mathbf{A}) \cap R(\mathbf{X}) \quad \text{(see Exercise 4.5.10)} \\ &= \dim N(\mathbf{A}) \cap R(\mathbf{B}). \quad \blacksquare \end{aligned}$$

The utility of (4.5.1) is mitigated by the fact that although $rank\,(\mathbf{A})$ and $rank\,(\mathbf{B})$ are frequently known or can be estimated, the term $\dim N(\mathbf{A}) \cap R(\mathbf{B})$ can be costly to obtain. In such cases (4.5.1) still provides us with useful upper and lower bounds for $rank\,(\mathbf{AB})$ that depend only on $rank\,(\mathbf{A})$ and $rank\,(\mathbf{B})$.

Bounds on the Rank of a Product

If $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then

- $$rank\,(\mathbf{AB}) \le \min\{rank\,(\mathbf{A}),\ rank\,(\mathbf{B})\}, \tag{4.5.2}$$
- $$rank\,(\mathbf{A}) + rank\,(\mathbf{B}) - n \le rank\,(\mathbf{AB}). \tag{4.5.3}$$

Proof. In words, (4.5.2) says that the rank of a product cannot exceed the rank of either factor. To prove $rank\,(\mathbf{AB}) \leq rank\,(\mathbf{B})$, use (4.5.1) and write

$$rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) \leq rank\,(\mathbf{B}).$$

This says that the rank of a product cannot exceed the rank of the right-hand factor. To show that $rank\,(\mathbf{AB}) \leq rank\,(\mathbf{A})$, remember that transposition does not alter rank, and use the reverse order law for transposes together with the previous statement to write

$$rank\,(\mathbf{AB}) = rank\,(\mathbf{AB})^T = rank\left(\mathbf{B}^T\mathbf{A}^T\right) \leq rank\left(\mathbf{A}^T\right) = rank\,(\mathbf{A}).$$

To prove (4.5.3), notice that $N\,(\mathbf{A}) \cap R\,(\mathbf{B}) \subseteq N\,(\mathbf{A})$, and recall from (4.4.5) that if $\mathcal{M}$ and $\mathcal{N}$ are spaces such that $\mathcal{M} \subseteq \mathcal{N}$, then $\dim \mathcal{M} \leq \dim \mathcal{N}$. Therefore,

$$\dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) \leq \dim N\,(\mathbf{A}) = n - rank\,(\mathbf{A}),$$

and the lower bound on $rank\,(\mathbf{AB})$ is obtained from (4.5.1) by writing

$$rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) \geq rank\,(\mathbf{B}) + rank\,(\mathbf{A}) - n. \quad \blacksquare$$

The products $\mathbf{A}^T\mathbf{A}$ and $\mathbf{AA}^T$ and their complex counterparts $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$ deserve special attention because they naturally appear in a wide variety of applications.

Products $\mathbf{A}^T\mathbf{A}$ and $\mathbf{AA}^T$

For $\mathbf{A} \in \Re^{m\times n}$, the following statements are true.

- $rank\left(\mathbf{A}^T\mathbf{A}\right) = rank\,(\mathbf{A}) = rank\left(\mathbf{AA}^T\right).$ (4.5.4)
- $R\left(\mathbf{A}^T\mathbf{A}\right) = R\left(\mathbf{A}^T\right)$ and $R\left(\mathbf{AA}^T\right) = R\,(\mathbf{A}).$ (4.5.5)
- $N\left(\mathbf{A}^T\mathbf{A}\right) = N\,(\mathbf{A})$ and $N\left(\mathbf{AA}^T\right) = N\left(\mathbf{A}^T\right).$ (4.5.6)

For $\mathbf{A} \in \mathcal{C}^{m\times n}$, the transpose operation $(\star)^T$ must be replaced by the conjugate transpose operation $(\star)^*$.

Proof. First observe that $N\left(\mathbf{A}^T\right) \cap R(\mathbf{A}) = \{\mathbf{0}\}$ because

$$\begin{aligned} \mathbf{x} \in N\left(\mathbf{A}^T\right) \cap R(\mathbf{A}) &\implies \mathbf{A}^T\mathbf{x} = \mathbf{0} \text{ and } \mathbf{x} = \mathbf{A}\mathbf{y} \text{ for some } \mathbf{y} \\ &\implies \mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}^T\mathbf{x} = 0 \implies \sum x_i^2 = 0 \\ &\implies \mathbf{x} = \mathbf{0}. \end{aligned}$$

Formula (4.5.1) for the rank of a product now guarantees that

$$rank\left(\mathbf{A}^T\mathbf{A}\right) = rank(\mathbf{A}) - \dim N\left(\mathbf{A}^T\right) \cap R(\mathbf{A}) = rank(\mathbf{A}),$$

which is half of (4.5.4)—the other half is obtained by reversing the roles of $\mathbf{A}$ and $\mathbf{A}^T$. To prove (4.5.5) and (4.5.6), use the facts $R(\mathbf{AB}) \subseteq R(\mathbf{A})$ and $N(\mathbf{B}) \subseteq N(\mathbf{AB})$ (see Exercise 4.2.12) to write $R\left(\mathbf{A}^T\mathbf{A}\right) \subseteq R\left(\mathbf{A}^T\right)$ and $N(\mathbf{A}) \subseteq N\left(\mathbf{A}^T\mathbf{A}\right)$. The first half of (4.5.5) and (4.5.6) now follows because

$$\dim R\left(\mathbf{A}^T\mathbf{A}\right) = rank\left(\mathbf{A}^T\mathbf{A}\right) = rank(\mathbf{A}) = rank\left(\mathbf{A}^T\right) = \dim R\left(\mathbf{A}^T\right),$$

$$\dim N(\mathbf{A}) = n - rank(\mathbf{A}) = n - rank\left(\mathbf{A}^T\mathbf{A}\right) = \dim N\left(\mathbf{A}^T\mathbf{A}\right).$$

Reverse the roles of $\mathbf{A}$ and $\mathbf{A}^T$ to get the second half of (4.5.5) and (4.5.6). ∎

To see why (4.5.4)—(4.5.6) might be important, consider an $m \times n$ system of equations $\mathbf{Ax} = \mathbf{b}$ that may or may not be consistent. Multiplying on the left-hand side by $\mathbf{A}^T$ produces the $n \times n$ system

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$$

called ***the associated system of normal equations***, which has some extremely interesting properties. First, notice that the normal equations are always consistent, regardless of whether or not the original system is consistent because (4.5.5) guarantees that $\mathbf{A}^T\mathbf{b} \in R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^T\mathbf{A}\right)$ (i.e., the right-hand side is in the range of the coefficient matrix), so (4.2.3) insures consistency. However, if $\mathbf{Ax} = \mathbf{b}$ happens to be consistent, then $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ have the same solution set because if $\mathbf{p}$ is a particular solution of the original system, then $\mathbf{Ap} = \mathbf{b}$ implies $\mathbf{A}^T\mathbf{Ap} = \mathbf{A}^T\mathbf{b}$ (i.e., $\mathbf{p}$ is also a particular solution of the normal equations), so the general solution of $\mathbf{Ax} = \mathbf{b}$ is $\mathcal{S} = \mathbf{p} + N(\mathbf{A})$, and the general solution of $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ is

$$\mathbf{p} + N\left(\mathbf{A}^T\mathbf{A}\right) = \mathbf{p} + N(\mathbf{A}) = \mathcal{S}.$$

Furthermore, if $\mathbf{Ax} = \mathbf{b}$ is consistent and has a unique solution, then the same is true for $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, and the unique solution common to both systems is

$$\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}. \tag{4.5.7}$$

This follows because a unique solution (to either system) exists if and only if $\mathbf{0} = N(\mathbf{A}) = N(\mathbf{A}^T\mathbf{A})$, and this insures $(\mathbf{A}^T\mathbf{A})_{n\times n}$ must be nonsingular (by (4.2.11)), so (4.5.7) is the unique solution to both systems. **Caution!** When $\mathbf{A}$ is not square, $\mathbf{A}^{-1}$ does not exist, and the reverse order law for inversion doesn't apply to $(\mathbf{A}^T\mathbf{A})^{-1}$, so (4.5.7) cannot be further simplified.

There is one outstanding question—what do the solutions of the normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ represent when the original system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is *not* consistent? The answer, which is of fundamental importance, will have to wait until §4.6, but let's summarize what has been said so far.

Normal Equations

- For an $m \times n$ system $\mathbf{A}\mathbf{x} = \mathbf{b}$, the associated system of ***normal equations*** is defined to be the $n \times n$ system $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$.
- $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ is always consistent, even when $\mathbf{A}\mathbf{x} = \mathbf{b}$ is not consistent.
- When $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent, its solution set agrees with that of $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. As discussed in §4.6, the normal equations provide least squares solutions to $\mathbf{A}\mathbf{x} = \mathbf{b}$ when $\mathbf{A}\mathbf{x} = \mathbf{b}$ is inconsistent.
- $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ has a unique solution if and only if $rank(\mathbf{A}) = n$, in which case the unique solution is $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
- When $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent and has a unique solution, then the same is true for $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$, and the unique solution to both systems is given by $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.

Example 4.5.1

Caution! Use of the product $\mathbf{A}^T\mathbf{A}$ or the normal equations is not recommended for numerical computation. Any sensitivity to small perturbations that is present in the underlying matrix $\mathbf{A}$ is magnified by forming the product $\mathbf{A}^T\mathbf{A}$. In other words, if $\mathbf{A}\mathbf{x} = \mathbf{b}$ is somewhat ill-conditioned, then the associated system of normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ will be ill-conditioned to an even greater extent, and the theoretical properties surrounding $\mathbf{A}^T\mathbf{A}$ and the normal equations may be lost in practical applications. For example, consider the nonsingular system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 1 & 2.01 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 9 \\ 3.01 \end{pmatrix}.$$

If Gaussian elimination with 3-digit floating-point arithmetic is used to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$, then the 3-digit solution is $(1,\ 1)$, and this agrees with the exact

solution. However if 3-digit arithmetic is used to form the associated system of normal equations, the result is

$$\begin{pmatrix} 10 & 20 \\ 20 & 40 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 30 \\ 60.1 \end{pmatrix}.$$

The 3-digit representation of $\mathbf{A}^T\mathbf{A}$ is singular, and the associated system of normal equations is inconsistent. For these reasons, the normal equations are often avoided in numerical computations. Nevertheless, the normal equations are an important theoretical idea that leads to practical tools of fundamental importance such as the method of least squares developed in §4.6 and §5.13.

Because the concept of rank is at the heart of our subject, it's important to understand rank from a variety of different viewpoints. The statement below is one more way to think about rank.[29]

Rank and the Largest Nonsingular Submatrix

The rank of a matrix $\mathbf{A}_{m\times n}$ is precisely the order of a maximal square nonsingular submatrix of $\mathbf{A}$. In other words, to say $rank\,(\mathbf{A}) = r$ means that there is at least one $r \times r$ nonsingular submatrix in $\mathbf{A}$, and there are no nonsingular submatrices of larger order.

Proof. First demonstrate that there exists an $r \times r$ nonsingular submatrix in $\mathbf{A}$, and then show there can be no nonsingular submatrix of larger order. Begin with the fact that there must be a maximal linearly independent set of r rows in $\mathbf{A}$ as well as a maximal independent set of r columns, and prove that the submatrix $\mathbf{M}_{r\times r}$ lying on the intersection of these r rows and r columns is nonsingular. The r independent rows can be permuted to the top, and the remaining rows can be annihilated using row operations, so

$$\mathbf{A} \overset{\text{row}}{\sim} \begin{pmatrix} \mathbf{U}_{r\times n} \\ \mathbf{0} \end{pmatrix}.$$

Now permute the r independent columns containing $\mathbf{M}$ to the left-hand side, and use column operations to annihilate the remaining columns to conclude that

$$\mathbf{A} \overset{\text{row}}{\sim} \begin{pmatrix} \mathbf{U}_{r\times n} \\ \mathbf{0} \end{pmatrix} \overset{\text{col}}{\sim} \begin{pmatrix} \mathbf{M}_{r\times r} & \mathbf{N} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \overset{\text{col}}{\sim} \begin{pmatrix} \mathbf{M}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

---

[29] This is the last characterization of rank presented in this text, but historically this was the essence of the first definition (p. 44) of rank given by Georg Frobenius (p. 662) in 1879.

Rank isn't changed by row or column operations, so $r = rank\,(\mathbf{A}) = rank\,(\mathbf{M})$, and thus $\mathbf{M}$ is nonsingular. Now suppose that $\mathbf{W}$ is any other nonsingular submatrix of $\mathbf{A}$, and let $\mathbf{P}$ and $\mathbf{Q}$ be permutation matrices such that $\mathbf{PAQ} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$. If

$$\mathbf{E} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{YW}^{-1} & \mathbf{I} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} \mathbf{I} & -\mathbf{W}^{-1}\mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \quad \text{and} \quad \mathbf{S} = \mathbf{Z} - \mathbf{YW}^{-1}\mathbf{X},$$

then

$$\mathbf{EPAQF} = \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix} \implies \mathbf{A} \sim \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}, \tag{4.5.8}$$

and hence $r = rank\,(\mathbf{A}) = rank\,(\mathbf{W}) + rank\,(\mathbf{S}) \geq rank\,(\mathbf{W})$ (recall Example 3.9.3). This guarantees that no nonsingular submatrix of $\mathbf{A}$ can have order greater than $r = rank\,(\mathbf{A})$. ∎

Example 4.5.2

Problem: Determine the rank of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \\ 3 & 6 & 1 \end{pmatrix}$.

Solution: $rank\,(\mathbf{A}) = 2$ because there is at least one 2×2 nonsingular submatrix (e.g., there is one lying on the intersection of rows 1 and 2 with columns 2 and 3), and there is no larger nonsingular submatrix (the entire matrix is singular). Notice that not all 2×2 matrices are nonsingular (e.g., consider the one lying on the intersection of rows 1 and 2 with columns 1 and 2).

Earlier in this section we saw that it is impossible to *increase* the rank by means of matrix multiplication—i.e., (4.5.2) says $rank\,(\mathbf{AE}) \leq rank\,(\mathbf{A})$. In a certain sense there is a dual statement for matrix addition that says that it is impossible to *decrease* the rank by means of a "small" matrix addition—i.e., $rank\,(\mathbf{A} + \mathbf{E}) \geq rank\,(\mathbf{A})$ whenever $\mathbf{E}$ has entries of small magnitude.

Small Perturbations Can't Reduce Rank

If $\mathbf{A}$ and $\mathbf{E}$ are $m \times n$ matrices such that $\mathbf{E}$ has entries of sufficiently small magnitude, then

$$rank\,(\mathbf{A} + \mathbf{E}) \geq rank\,(\mathbf{A}). \tag{4.5.9}$$

The term "sufficiently small" is further clarified in Exercise 5.12.4.

Proof. Suppose $rank\,(\mathbf{A}) = r$, and let $\mathbf{P}$ and $\mathbf{Q}$ be nonsingular matrices that reduce $\mathbf{A}$ to rank normal form—i.e., $\mathbf{PAQ} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. If $\mathbf{P}$ and $\mathbf{Q}$ are applied to $\mathbf{E}$ to form $\mathbf{PEQ} = \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}$, where $\mathbf{E}_{11}$ is $r \times r$, then

$$\mathbf{P}(\mathbf{A}+\mathbf{E})\mathbf{Q} = \begin{pmatrix} \mathbf{I}_r + \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}. \tag{4.5.10}$$

If the magnitude of the entries in $\mathbf{E}$ are small enough to insure that $\mathbf{E}_{11}^k \to \mathbf{0}$ as $k \to \infty$, then the discussion of the Neumann series on p. 126 insures that $\mathbf{I} + \mathbf{E}_{11}$ is nonsingular. (Exercise 4.5.14 gives another condition on the size of $\mathbf{E}_{11}$ to insure this.) This allows the right-hand side of (4.5.10) to be further reduced by writing

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{E}_{21}(\mathbf{I}+\mathbf{E}_{11})^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I}+\mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -(\mathbf{I}+\mathbf{E}_{11})^{-1}\mathbf{E}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{I}-\mathbf{E}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix},$$

where $\mathbf{S} = \mathbf{E}_{22} - \mathbf{E}_{21}\left(\mathbf{I}+\mathbf{E}_{11}\right)^{-1}\mathbf{E}_{12}$. In other words,

$$\mathbf{A}+\mathbf{E} \sim \begin{pmatrix} \mathbf{I}-\mathbf{E}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix},$$

and therefore

$$\begin{aligned} rank\,(\mathbf{A}+\mathbf{E}) &= rank\,(\mathbf{I}_r + \mathbf{E}_{11}) + rank\,(\mathbf{S}) \quad \text{(recall Example 3.9.3)} \\ &= rank\,(\mathbf{A}) + rank\,(\mathbf{S}) \\ &\geq rank\,(\mathbf{A}). \quad \blacksquare \end{aligned} \tag{4.5.11}$$

Example 4.5.3

A Pitfall in Solving Singular Systems. Solving $\mathbf{Ax} = \mathbf{b}$ with floating-point arithmetic produces the exact solution of a perturbed system whose coefficient matrix is $\mathbf{A}+\mathbf{E}$. If $\mathbf{A}$ is nonsingular, and if we are using a stable algorithm (an algorithm that insures that the entries in $\mathbf{E}$ have small magnitudes), then (4.5.9) guarantees that we are finding the exact solution to a nearby system that is also nonsingular. On the other hand, if $\mathbf{A}$ is singular, then perturbations of even the slightest magnitude can increase the rank, thereby producing a system with fewer free variables than the original system theoretically demands, so even a stable algorithm can result in a significant loss of information. But what are the chances that this will actually occur in practice? To answer this, recall from (4.5.11) that

$$rank\,(\mathbf{A}+\mathbf{E}) = rank\,(\mathbf{A}) + rank\,(\mathbf{S}), \quad \text{where} \quad \mathbf{S} = \mathbf{E}_{22} - \mathbf{E}_{21}\left(\mathbf{I}+\mathbf{E}_{11}\right)^{-1}\mathbf{E}_{12}.$$

If the rank is not to jump, then the perturbation $\mathbf{E}$ must be such that $\mathbf{S} = \mathbf{0}$, which is equivalent to saying $\mathbf{E}_{22} = \mathbf{E}_{21}\left(\mathbf{I} + \mathbf{E}_{11}\right)^{-1}\mathbf{E}_{12}$. Clearly, this requires the existence of a very specific (and quite special) relationship among the entries of $\mathbf{E}$, and a random perturbation will almost never produce such a relationship. Although rounding errors cannot be considered to be truly random, they are random enough so as to make the possibility that $\mathbf{S} = \mathbf{0}$ very unlikely. Consequently, when $\mathbf{A}$ is singular, the small perturbation $\mathbf{E}$ due to roundoff makes the possibility that $rank\,(\mathbf{A} + \mathbf{E}) > rank\,(\mathbf{A})$ very likely. The moral is to avoid floating-point solutions of singular systems. Singular problems can often be distilled down to a nonsingular core or to nonsingular pieces, and these are the components you should be dealing with.

Since no more significant characterizations of rank will be given, it is appropriate to conclude this section with a summary of all of the different ways we have developed to say "rank."

Summary of Rank

For $\mathbf{A} \in \Re^{m\times n}$, each of the following statements is true.

- $rank\,(\mathbf{A}) =$ The number of nonzero rows in any row echelon form that is row equivalent to $\mathbf{A}$.
- $rank\,(\mathbf{A}) =$ The number of pivots obtained in reducing $\mathbf{A}$ to a row echelon form with row operations.
- $rank\,(\mathbf{A}) =$ The number of basic columns in $\mathbf{A}$ (as well as the number of basic columns in any matrix that is row equivalent to $\mathbf{A}$).
- $rank\,(\mathbf{A}) =$ The number of independent columns in $\mathbf{A}$—i.e., the size of a maximal independent set of columns from $\mathbf{A}$.
- $rank\,(\mathbf{A}) =$ The number of independent rows in $\mathbf{A}$—i.e., the size of a maximal independent set of rows from $\mathbf{A}$.
- $rank\,(\mathbf{A}) = \dim R\,(\mathbf{A})$.
- $rank\,(\mathbf{A}) = \dim R\left(\mathbf{A}^T\right)$.
- $rank\,(\mathbf{A}) = n - \dim N\,(\mathbf{A})$.
- $rank\,(\mathbf{A}) = m - \dim N\left(\mathbf{A}^T\right)$.
- $rank\,(\mathbf{A}) =$ The size of the largest nonsingular submatrix in $\mathbf{A}$.

For $\mathbf{A} \in \mathcal{C}^{m\times n}$, replace $(\star)^T$ with $(\star)^*$.

Exercises for section 4.5

4.5.1. Verify that $rank\left(\mathbf{A}^T\mathbf{A}\right) = rank\left(\mathbf{A}\right) = rank\left(\mathbf{A}\mathbf{A}^T\right)$ for

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

4.5.2. Determine $\dim N\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)$ for

$$\mathbf{A} = \begin{pmatrix} -2 & 1 & 1 \\ -4 & 2 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 3 & 1 & -4 \\ -1 & -3 & 1 & 0 \\ 2 & 6 & 2 & -8 \end{pmatrix}.$$

4.5.3. For the matrices given in Exercise 4.5.2, use the procedure described on p. 211 to determine a basis for $N\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)$.

4.5.4. If $\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_k$ is a product of square matrices such that some $\mathbf{A}_i$ is singular, explain why the entire product must be singular.

4.5.5. For $\mathbf{A} \in \Re^{m\times n}$, explain why $\mathbf{A}^T\mathbf{A} = \mathbf{0}$ implies $\mathbf{A} = \mathbf{0}$.

4.5.6. Find $rank\left(\mathbf{A}\right)$ and all nonsingular submatrices of maximal order in

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 1 \\ 4 & -2 & 1 \\ 8 & -4 & 1 \end{pmatrix}.$$

4.5.7. Is it possible that $rank\left(\mathbf{AB}\right) < rank\left(\mathbf{A}\right)$ and $rank\left(\mathbf{AB}\right) < rank\left(\mathbf{B}\right)$ for the same pair of matrices?

4.5.8. Is $rank\left(\mathbf{AB}\right) = rank\left(\mathbf{BA}\right)$ when both products are defined? Why?

4.5.9. Explain why $rank\left(\mathbf{AB}\right) = rank\left(\mathbf{A}\right) - \dim N\left(\mathbf{B}^T\right) \cap R\left(\mathbf{A}^T\right)$.

4.5.10. Explain why $\dim N\left(\mathbf{A}_{m\times n}\mathbf{B}_{n\times p}\right) = \dim N\left(\mathbf{B}\right) + \dim R\left(\mathbf{B}\right) \cap N\left(\mathbf{A}\right)$.

4.5.11. ***Sylvester's law of nullity,*** given by James J. Sylvester in 1884, states that for *square matrices* $\mathbf{A}$ and $\mathbf{B}$,

$$\max\{\nu(\mathbf{A}),\ \nu(\mathbf{B})\} \le \nu(\mathbf{AB}) \le \nu(\mathbf{A}) + \nu(\mathbf{B}),$$

where $\nu(\star) = \dim N(\star)$ denotes the nullity.

(a) Establish the validity of Sylvester's law.

(b) Show Sylvester's law is not valid for rectangular matrices because $\nu(\mathbf{A}) > \nu(\mathbf{AB})$ is possible. Is $\nu(\mathbf{B}) > \nu(\mathbf{AB})$ possible?

4.5.12. For matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times p}$, prove each of the following statements:

(a) $rank\,(\mathbf{AB}) = rank\,(\mathbf{A})$ and $R(\mathbf{AB}) = R(\mathbf{A})$ if $rank\,(\mathbf{B}) = n$.

(b) $rank\,(\mathbf{AB}) = rank\,(\mathbf{B})$ and $N(\mathbf{AB}) = N(\mathbf{B})$ if $rank\,(\mathbf{A}) = n$.

4.5.13. Perform the following calculations using the matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 1 & 2.01 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 1.01 \end{pmatrix}.$$

(a) Find $rank\,(\mathbf{A})$, and solve $\mathbf{Ax} = \mathbf{b}$ using exact arithmetic.

(b) Find $rank\,\left(\mathbf{A}^T\mathbf{A}\right)$, and solve $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ exactly.

(c) Find $rank\,(\mathbf{A})$, and solve $\mathbf{Ax} = \mathbf{b}$ with 3-digit arithmetic.

(d) Find $\mathbf{A}^T\mathbf{A}$, $\mathbf{A}^T\mathbf{b}$, and the solution of $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ with 3-digit arithmetic.

4.5.14. Prove that if the entries of $\mathbf{F}_{r\times r}$ satisfy $\sum_{j=1}^{r} |f_{ij}| < 1$ for each i (i.e., each absolute row sum < 1), then $\mathbf{I} + \mathbf{F}$ is nonsingular. **Hint:** Use the triangle inequality for scalars $|\alpha+\beta| \le |\alpha|+|\beta|$ to show $N(\mathbf{I}+\mathbf{F}) = \mathbf{0}$.

4.5.15. If $\mathbf{A} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$, where $rank\,(\mathbf{A}) = r = rank\,(\mathbf{W}_{r\times r})$, show that there are matrices $\mathbf{B}$ and $\mathbf{C}$ such that

$$\mathbf{A} = \begin{pmatrix} \mathbf{W} & \mathbf{WC} \\ \mathbf{BW} & \mathbf{BWC} \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{B} \end{pmatrix} \mathbf{W}\left(\mathbf{I} \mid \mathbf{C}\right).$$

4.5.16. For a convergent sequence $\{\mathbf{A}_k\}_{k=1}^{\infty}$ of matrices, let $\mathbf{A} = \lim_{k\to\infty} \mathbf{A}_k$.

(a) Prove that if each $\mathbf{A}_k$ is singular, then $\mathbf{A}$ is singular.

(b) If each $\mathbf{A}_k$ is nonsingular, must $\mathbf{A}$ be nonsingular? Why?

4.5.17. The Frobenius Inequality. Establish the validity of Frobenius's 1911 result that states that if $\mathbf{ABC}$ exists, then

$$rank\,(\mathbf{AB}) + rank\,(\mathbf{BC}) \leq rank\,(\mathbf{B}) + rank\,(\mathbf{ABC}).$$

Hint: If $\mathcal{M} = R\,(\mathbf{BC}) \cap N\,(\mathbf{A})$ and $\mathcal{N} = R\,(\mathbf{B}) \cap N\,(\mathbf{A})$, then $\mathcal{M} \subseteq \mathcal{N}$.

4.5.18. If $\mathbf{A}$ is $n \times n$, prove that the following statements are equivalent:
(a) $N\,(\mathbf{A}) = N\left(\mathbf{A}^2\right)$.
(b) $R\,(\mathbf{A}) = R\left(\mathbf{A}^2\right)$.
(c) $R\,(\mathbf{A}) \cap N\,(\mathbf{A}) = \{\mathbf{0}\}$.

4.5.19. Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ matrices such that $\mathbf{A} = \mathbf{A}^2$, $\mathbf{B} = \mathbf{B}^2$, and $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$.
(a) Prove that $rank\,(\mathbf{A} + \mathbf{B}) = rank\,(\mathbf{A}) + rank\,(\mathbf{B})$. **Hint:** Consider $\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{A} + \mathbf{B})(\mathbf{A} \mid \mathbf{B})$.
(b) Prove that $rank\,(\mathbf{A}) + rank\,(\mathbf{I} - \mathbf{A}) = n$.

4.5.20. Moore–Penrose Inverse. For $\mathbf{A} \in \Re^{m \times n}$ such that $rank\,(\mathbf{A}) = r$, let $\mathbf{A} = \mathbf{BC}$ be the full rank factorization of $\mathbf{A}$ in which $\mathbf{B}_{m \times r}$ is the matrix of basic columns from $\mathbf{A}$ and $\mathbf{C}_{r \times n}$ is the matrix of nonzero rows from $\mathbf{E_A}$ (see Exercise 3.9.8). The matrix defined by

$$\mathbf{A}^\dagger = \mathbf{C}^T \left(\mathbf{B}^T \mathbf{A} \mathbf{C}^T\right)^{-1} \mathbf{B}^T$$

is called the ***Moore–Penrose***[30] ***inverse*** of $\mathbf{A}$. Some authors refer to $\mathbf{A}^\dagger$ as the *pseudoinverse* or the *generalized inverse* of $\mathbf{A}$. A more elegant treatment is given on p. 423, but it's worthwhile to introduce the idea here so that it can be used and viewed from different perspectives.
(a) Explain why the matrix $\mathbf{B}^T \mathbf{A} \mathbf{C}^T$ is nonsingular.
(b) Verify that $\mathbf{x} = \mathbf{A}^\dagger \mathbf{b}$ solves the normal equations $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b}$ (as well as $\mathbf{Ax} = \mathbf{b}$ when it is consistent).
(c) Show that the general solution for $\mathbf{A}^T \mathbf{A} \mathbf{x} = \mathbf{A}^T \mathbf{b}$ (as well as $\mathbf{Ax} = \mathbf{b}$ when it is consistent) can be described as

$$\mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + \left(\mathbf{I} - \mathbf{A}^\dagger \mathbf{A}\right) \mathbf{h},$$

[30] This is in honor of Eliakim H. Moore (1862–1932) and Roger Penrose (a famous contemporary English mathematical physicist). Each formulated a concept of generalized matrix inversion—Moore's work was published in 1922, and Penrose's work appeared in 1955. E. H. Moore is considered by many to be America's first great mathematician.

where $\mathbf{h}$ is a "free variable" vector in $\Re^{n\times 1}$.
Hint: Verify $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$, and then show $R\left(\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A}\right) = N\left(\mathbf{A}\right)$.

(d) If $rank\left(\mathbf{A}\right) = n$, explain why $\mathbf{A}^{\dagger} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$.

(e) If $\mathbf{A}$ is square and nonsingular, explain why $\mathbf{A}^{\dagger} = \mathbf{A}^{-1}$.

(f) Verify that $\mathbf{A}^{\dagger} = \mathbf{C}^T\left(\mathbf{B}^T\mathbf{A}\mathbf{C}^T\right)^{-1}\mathbf{B}^T$ satisfies the Penrose equations:

$$\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}, \qquad \left(\mathbf{A}\mathbf{A}^{\dagger}\right)^T = \mathbf{A}\mathbf{A}^{\dagger},$$

$$\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}, \qquad \left(\mathbf{A}^{\dagger}\mathbf{A}\right)^T = \mathbf{A}^{\dagger}\mathbf{A}.$$

Penrose originally defined $\mathbf{A}^{\dagger}$ to be the unique solution to these four equations.

4.6 CLASSICAL LEAST SQUARES

The following problem arises in almost all areas where mathematics is applied. At discrete points t_i (often points in time), observations b_i of some phenomenon are made, and the results are recorded as a set of ordered pairs

$$\mathcal{D} = \{(t_1, b_1), \quad (t_2, b_2), \ \ldots, \ (t_m, b_m)\}.$$

On the basis of these observations, the problem is to make estimations or predictions at points (times) $\hat{t}$ that are between or beyond the observation points t_i. A standard approach is to find the equation of a curve $y = f(t)$ that closely fits the points in $\mathcal{D}$ so that the phenomenon can be estimated at any nonobservation point $\hat{t}$ with the value $\hat{y} = f(\hat{t})$.

Let's begin by fitting a straight line to the points in $\mathcal{D}$. Once this is understood, it will be relatively easy to see how to fit the data with curved lines.

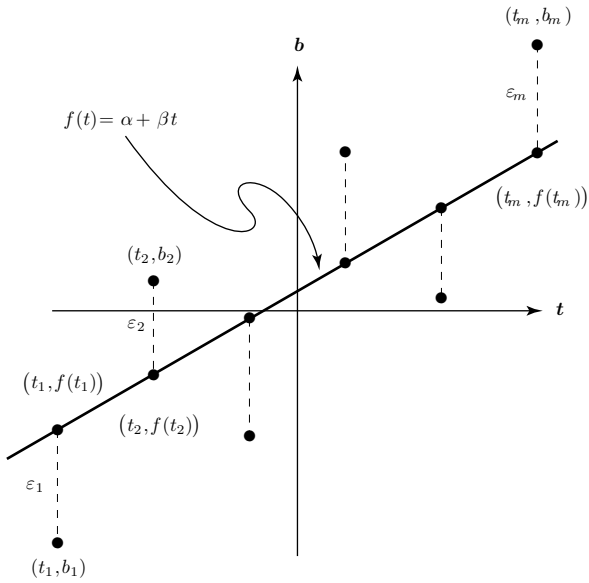


FIGURE 4.6.1

The strategy is to determine the coefficients α and β in the equation of the line $f(t) = \alpha + \beta t$ that best fits the points (t_i, b_i) in the sense that the sum of the squares of the vertical[31] errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ indicated in Figure 4.6.1 is

[31] We consider only vertical errors because there is a tacit assumption that only the observations b_i are subject to error or variation. The t_i's are assumed to be errorless constants—think of them as being exact points in time (as they often are). If the t_i's are also subject to variation, then horizontal as well as vertical errors have to be considered in Figure 4.6.1, and a more complicated theory known as *total least squares* (not considered in this text) emerges. The least squares line $\mathcal{L}$ obtained by minimizing only vertical deviations will not be the closest line to points in $\mathcal{D}$ in terms of perpendicular distance, but $\mathcal{L}$ is the best line for the purpose of linear estimation—see §5.14 (p. 446).

minimal. The distance from (t_i, b_i) to a line $f(t) = \alpha + \beta t$ is

$$\varepsilon_i = |f(t_i) - b_i| = |\alpha + \beta t_i - b_i|,$$

so that the objective is to find values for α and β such that

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \quad \text{is minimal.}$$

Minimization techniques from calculus tell us that the minimum value must occur at a solution to the two equations

$$0 = \frac{\partial \left(\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \right)}{\partial \alpha} = 2 \sum_{i=1}^{m} (\alpha + \beta t_i - b_i),$$

$$0 = \frac{\partial \left(\sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 \right)}{\partial \beta} = 2 \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)\, t_i.$$

Rearranging terms produces two equations in the two unknowns α and β

$$\begin{aligned} \left(\sum_{i=1}^{m} 1 \right) \alpha + \left(\sum_{i=1}^{m} t_i \right) \beta &= \sum_{i=1}^{m} b_i, \\ \left(\sum_{i=1}^{m} t_i \right) \alpha + \left(\sum_{i=1}^{m} t_i^2 \right) \beta &= \sum_{i=1}^{m} t_i b_i. \end{aligned} \tag{4.6.1}$$

By setting

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

we see that the two equations (4.6.1) have the matrix form $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. In other words, (4.6.1) is the system of normal equations associated with the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ (see p. 213). The t_i's are assumed to be distinct numbers, so $rank\,(\mathbf{A}) = 2$, and (4.5.7) insures that the normal equations have a unique solution given by

$$\begin{aligned} \mathbf{x} &= \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b} \\ &= \frac{1}{m\sum t_i^2 - \left(\sum t_i\right)^2} \begin{pmatrix} \sum t_i^2 & -\sum t_i \\ -\sum t_i & m \end{pmatrix} \begin{pmatrix} \sum b_i \\ \sum t_i b_i \end{pmatrix} \\ &= \frac{1}{m\sum t_i^2 - \left(\sum t_i\right)^2} \begin{pmatrix} \sum t_i^2 \sum b_i - \sum t_i \sum t_i b_i \\ m\sum t_i b_i - \sum t_i \sum b_i \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \end{aligned}$$

Finally, notice that the total sum of squares of the errors is given by

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (\alpha + \beta t_i - b_i)^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Example 4.6.1

Problem: A small company has been in business for four years and has recorded annual sales (in tens of thousands of dollars) as follows.

Year	1	2	3	4
Sales	23	27	30	34

When this data is plotted as shown in Figure 4.6.2, we see that although the points do not exactly lie on a straight line, there nevertheless appears to be a linear trend. Predict the sales for any future year if this trend continues.

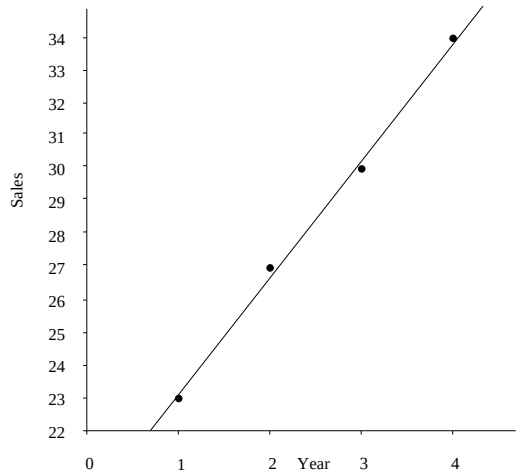


FIGURE 4.6.2

Solution: Determine the line $f(t) = \alpha + \beta t$ that best fits the data in the sense of least squares. If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 23 \\ 27 \\ 30 \\ 34 \end{pmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

then the previous discussion guarantees that $\mathbf{x}$ is the solution of the normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. That is,

$$\begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 114 \\ 303 \end{pmatrix}.$$

The solution is easily found to be $\alpha = 19.5$ and $\beta = 3.6$, so we predict that the sales in year t will be $f(t) = 19.5 + 3.6t$. For example, the estimated sales for year five is \$375,000. To get a feel for how close the least squares line comes to

passing through the data points, let $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, and compute the sum of the squares of the errors to be

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = .2.$$

General Least Squares Problem

For $\mathbf{A} \in \Re^{m \times n}$ and $\mathbf{b} \in \Re^m$, let $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{x}) = \mathbf{Ax} - \mathbf{b}$. The general least squares problem is to find a vector $\mathbf{x}$ that minimizes the quantity

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}).$$

Any vector that provides a minimum value for this expression is called a ***least squares solution.***

- The set of all least squares solutions is precisely the set of solutions to the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.
- There is a unique least squares solution if and only if $rank\,(\mathbf{A}) = n$, in which case it is given by $\mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$.
- If $\mathbf{Ax} = \mathbf{b}$ is consistent, then the solution set for $\mathbf{Ax} = \mathbf{b}$ is the same as the set of least squares solutions.

Proof.[32] First prove that if $\mathbf{x}$ minimizes $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$, then $\mathbf{x}$ must satisfy the normal equations. Begin by using $\mathbf{x}^T\mathbf{A}^T\mathbf{b} = \mathbf{b}^T\mathbf{Ax}$ (scalars are symmetric) to write

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b}. \tag{4.6.2}$$

To determine vectors $\mathbf{x}$ that minimize the expression (4.6.2), we will again use minimization techniques from calculus and differentiate the function

$$f(x_1, x_2, \ldots, x_n) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{x}^T\mathbf{A}^T\mathbf{b} + \mathbf{b}^T\mathbf{b} \tag{4.6.3}$$

with respect to each x_i. Differentiating matrix functions is similar to differentiating scalar functions (see Exercise 3.5.9) in the sense that if $\mathbf{U} = [u_{ij}]$, then

$$\left[\frac{\partial \mathbf{U}}{\partial x}\right]_{ij} = \frac{\partial u_{ij}}{\partial x}, \quad \frac{\partial[\mathbf{U} + \mathbf{V}]}{\partial x} = \frac{\partial \mathbf{U}}{\partial x} + \frac{\partial \mathbf{V}}{\partial x}, \quad \text{and} \quad \frac{\partial[\mathbf{UV}]}{\partial x} = \frac{\partial \mathbf{U}}{\partial x}\mathbf{V} + \mathbf{U}\frac{\partial \mathbf{V}}{\partial x}.$$

[32] A more modern development not relying on calculus is given in §5.13 on p. 437, but the more traditional approach is given here because it's worthwhile to view least squares from both perspectives.

Applying these rules to the function in (4.6.3) produces

$$\frac{\partial f}{\partial x_i} = \frac{\partial \mathbf{x}^T}{\partial x_i}\mathbf{A}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\frac{\partial \mathbf{x}}{\partial x_i} - 2\frac{\partial \mathbf{x}^T}{\partial x_i}\mathbf{A}^T\mathbf{b}.$$

Since $\partial\mathbf{x}/\partial x_i = \mathbf{e}_i$ (the ith unit vector), we have

$$\frac{\partial f}{\partial x_i} = \mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{e}_i - 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{b} = 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{e}_i^T\mathbf{A}^T\mathbf{b}.$$

Using $\mathbf{e}_i^T\mathbf{A}^T = \left(\mathbf{A}^T\right)_{i*}$ and setting $\partial f/\partial x_i = 0$ produces the n equations

$$\left(\mathbf{A}^T\right)_{i*}\mathbf{A}\mathbf{x} = \left(\mathbf{A}^T\right)_{i*}\mathbf{b} \quad \text{for } i = 1, 2, \ldots, n,$$

which can be written as the single matrix equation $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Calculus guarantees that the minimum value of f occurs at *some* solution of this system. But this is not enough—we want to know that *every* solution of $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$ is a least squares solution. So we must show that the function f in (4.6.3) attains its minimum value at each solution to $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. Observe that if $\mathbf{z}$ is a solution to the normal equations, then $f(\mathbf{z}) = \mathbf{b}^T\mathbf{b} - \mathbf{z}^T\mathbf{A}^T\mathbf{b}$. For any other $\mathbf{y} \in \Re^{n\times 1}$, let $\mathbf{u} = \mathbf{y} - \mathbf{z}$, so $\mathbf{y} = \mathbf{z} + \mathbf{u}$, and observe that

$$f(\mathbf{y}) = f(\mathbf{z}) + \mathbf{v}^T\mathbf{v}, \quad \text{where} \quad \mathbf{v} = \mathbf{A}\mathbf{u}.$$

Since $\mathbf{v}^T\mathbf{v} = \sum_i \mathbf{v}_i^2 \geq 0$, it follows that $f(\mathbf{z}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in \Re^{n\times 1}$, and thus f attains its minimum value at each solution of the normal equations. The remaining statements in the theorem follow from the properties established on p. 213. ∎

The classical least squares problem discussed at the beginning of this section and illustrated in Example 4.6.1 is part of a broader topic known as *linear regression,* which is the study of situations where attempts are made to express one variable y as a linear combination of other variables $t_1, t_2, \ldots, t_n$. In practice, hypothesizing that y is linearly related to $t_1, t_2, \ldots, t_n$ means that one assumes the existence of a set of constants $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$ (called *parameters*) such that

$$y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n + \varepsilon,$$

where ε is a "random function" whose values "average out" to zero in some sense. Practical problems almost always involve more variables than we wish to consider, but it is frequently fair to assume that the effect of variables of lesser significance will indeed "average out" to zero. The random function ε accounts for this assumption. In other words, a linear hypothesis is the supposition that the expected (or mean) value of y at each point where the phenomenon can be observed is given by a linear equation

$$E(y) = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n.$$

To help seat these ideas, consider the problem of predicting the amount of weight that a pint of ice cream loses when it is stored at very low temperatures. There are many factors that may contribute to weight loss—e.g., storage temperature, storage time, humidity, atmospheric pressure, butterfat content, the amount of corn syrup, the amounts of various gums (guar gum, carob bean gum, locust bean gum, cellulose gum), and the never-ending list of other additives and preservatives. It is reasonable to believe that storage time and temperature are the primary factors, so to predict weight loss we will make a linear hypothesis of the form

$$y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + \varepsilon,$$

where $y =$ weight loss (grams), $t_1 =$ storage time (weeks), $t_2 =$ storage temperature (^oF), and ε is a random function to account for all other factors. The assumption is that all other factors "average out" to zero, so the expected (or mean) weight loss at each point (t_1, t_2) is

$$E(y) = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2. \tag{4.6.4}$$

Suppose that we conduct an experiment in which values for weight loss are measured for various values of storage time and temperature as shown below.

Time (weeks)	1	1	1	2	2	2	3	3	3
Temp (^oF)	−10	−5	0	−10	−5	0	−10	−5	0
Loss (grams)	.15	.18	.20	.17	.19	.22	.20	.23	.25

If

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & -10 \\ 1 & 1 & -5 \\ 1 & 1 & 0 \\ 1 & 2 & -10 \\ 1 & 2 & -5 \\ 1 & 2 & 0 \\ 1 & 3 & -10 \\ 1 & 3 & -5 \\ 1 & 3 & 0 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} .15 \\ .18 \\ .20 \\ .17 \\ .19 \\ .22 \\ .20 \\ .23 \\ .25 \end{pmatrix},$$

and if we were lucky enough to exactly observe the mean weight loss each time (i.e., if $\mathbf{b}_i = E(y_i)$), then equation (4.6.4) would insure that $\mathbf{Ax} = \mathbf{b}$ is a consistent system, so we could solve for the unknown parameters $\alpha_0, \alpha_1,$ and α_2. However, it is virtually impossible to observe the *exact* value of the mean weight loss for a given storage time and temperature, and almost certainly the system defined by $\mathbf{Ax} = \mathbf{b}$ will be inconsistent—especially when the number of observations greatly exceeds the number of parameters. Since we can't solve $\mathbf{Ax} = \mathbf{b}$ to find exact values for the α_i 's, the best we can hope for is a set of "good estimates" for these parameters.

The famous Gauss–Markov theorem (developed on p. 448) states that under certain reasonable assumptions concerning the random error function ε, the "best" estimates for the α_i's are obtained by minimizing the sum of squares $(\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b})$. In other words, the least squares estimates are the "best" way to estimate the α_i's.

Returning to our ice cream example, it can be verified that $\mathbf{b} \notin R(\mathbf{A})$, so, as expected, the system $\mathbf{Ax}=\mathbf{b}$ is not consistent, and we cannot determine exact values for α_0, α_1, and α_2. The best we can do is to determine least squares estimates for the α_i's by solving the associated normal equations $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$, which in this example are

$$\begin{pmatrix} 9 & 18 & -45 \\ 18 & 42 & -90 \\ -45 & -90 & 375 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1.79 \\ 3.73 \\ -8.2 \end{pmatrix}.$$

The solution is

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} .174 \\ .025 \\ .005 \end{pmatrix},$$

and the estimating equation for mean weight loss becomes

$$\hat{y} = .174 + .025t_1 + .005t_2.$$

For example, the mean weight loss of a pint of ice cream that is stored for nine weeks at a temperature of -35^oF is estimated to be

$$\hat{y} = .174 + .025(9) + .005(-35) = .224 \text{ grams.}$$

Example 4.6.2

Least Squares Curve Fitting Problem: Find a polynomial

$$p(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_{n-1}t^{n-1}$$

with a specified degree that comes as close as possible in the sense of least squares to passing through a set of data points

$$\mathcal{D} = \{(t_1, b_1), \quad (t_2, b_2), \ \ldots, \ (t_m, b_m)\},$$

where the t_i's are distinct numbers, and $n \leq m$.

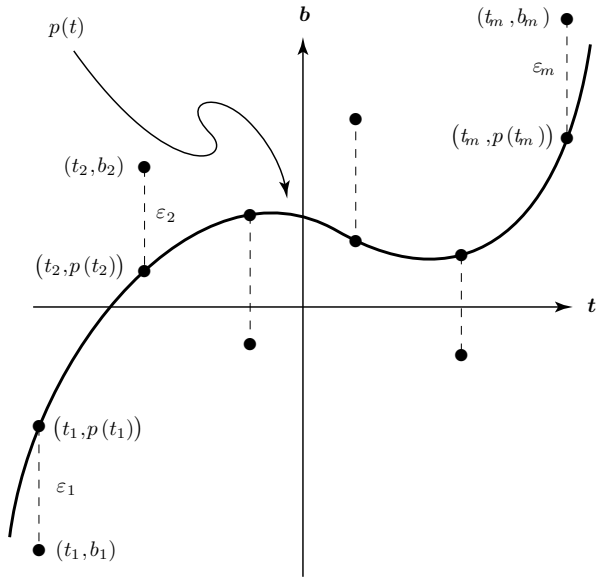


FIGURE 4.6.3

Solution: For the ε_i 's indicated in Figure 4.6.3, the objective is to minimize the sum of squares

$$\sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} \big(p(t_i) - b_i\big)^2 = (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}),$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^{n-1} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

In other words, the least squares polynomial of degree $n-1$ is obtained from the least squares solution associated with the system $\mathbf{Ax} = \mathbf{b}$. Furthermore, this least squares polynomial is unique because $\mathbf{A}_{m\times n}$ is the Vandermonde matrix of Example 4.3.4 with $n \le m$, so $rank\,(\mathbf{A}) = n$, and $\mathbf{Ax} = \mathbf{b}$ has a unique least squares solution given by $\mathbf{x} = \big(\mathbf{A}^T\mathbf{A}\big)^{-1}\mathbf{A}^T\mathbf{b}$.

Note: We know from Example 4.3.5 on p. 186 that the Lagrange interpolation polynomial $\ell(t)$ of degree $m-1$ will *exactly* fit the data—i.e., it passes through each point in $\mathcal{D}$. So why would one want to settle for a least squares fit when an exact fit is possible? One answer stems from the fact that in practical work the observations b_i are rarely exact due to small errors arising from imprecise

measurements or from simplifying assumptions. For this reason, it is the *trend* of the observations that needs to be fitted and not the observations themselves. To hit the data points, the interpolation polynomial $\ell(t)$ is usually forced to oscillate between or beyond the data points, and as m becomes larger the oscillations can become more pronounced. Consequently, $\ell(t)$ is generally not useful in making estimations concerning the trend of the observations—Example 4.6.3 drives this point home. In addition to exactly hitting a prescribed set of data points, an interpolation polynomial called the *Hermite polynomial* (p. 607) can be constructed to have specified derivatives at each data point. While this helps, it still is not as good as least squares for making estimations on the basis of observations.

Example 4.6.3

A missile is fired from enemy territory, and its position in flight is observed by radar tracking devices at the following positions.

Position down range (miles)	0	250	500	750	1000
Height (miles)	0	8	15	19	20

Suppose our intelligence sources indicate that enemy missiles are programmed to follow a parabolic flight path—a fact that seems to be consistent with the diagram obtained by plotting the observations on the coordinate system shown in Figure 4.6.4.

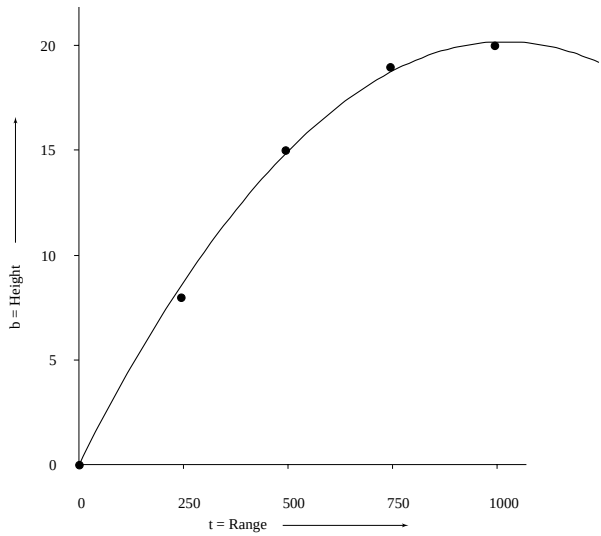


FIGURE 4.6.4

Problem: Predict how far down range the missile will land.

Solution: Determine the parabola $f(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ that best fits the observed data in the least squares sense. Then estimate where the missile will land by determining the roots of f (i.e., determine where the parabola crosses the horizontal axis). As it stands, the problem will involve numbers having relatively large magnitudes in conjunction with relatively small ones. Consequently, it is better to first scale the data by considering one unit to be 1000 miles. If

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .25 & .0625 \\ 1 & .5 & .25 \\ 1 & .75 & .5625 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 0 \\ .008 \\ .015 \\ .019 \\ .02 \end{pmatrix},$$

and if $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, then the object is to find a least squares solution $\mathbf{x}$ that minimizes

$$\sum_{i=1}^{5} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}).$$

We know that such a least squares solution is given by the solution to the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, which in this case is

$$\begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828125 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} .062 \\ .04375 \\ .0349375 \end{pmatrix}.$$

The solution (rounded to four significant digits) is

$$\mathbf{x} = \begin{pmatrix} -2.286 \times 10^{-4} \\ 3.983 \times 10^{-2} \\ -1.943 \times 10^{-2} \end{pmatrix},$$

and the least squares parabola is

$$f(t) = -.0002286 + .03983t - .01943t^2.$$

To estimate where the missile will land, determine where this parabola crosses the horizontal axis by applying the quadratic formula to find the roots of $f(t)$ to be $t = .005755$ and $t = 2.044$. Therefore, we estimate that the missile will land 2044 miles down range. The sum of the squares of the errors associated with the least squares solution is

$$\sum_{i=1}^{5} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = 4.571 \times 10^{-7}.$$

Least Squares vs. Lagrange Interpolation. Instead of using least squares, fit the observations exactly with the fourth-degree Lagrange interpolation polynomial

$$\ell(t) = \frac{11}{375}t + \frac{17}{750000}t^2 - \frac{1}{18750000}t^3 + \frac{1}{46875000000}t^4$$

described in Example 4.3.5 on p. 186 (you can verify that $\ell(t_i) = b_i$ for each observation). As the graph in Figure 4.6.5 indicates, $\ell(t)$ has only one real nonnegative root, so it is worthless for predicting where the missile will land. This is characteristic of Lagrange interpolation.

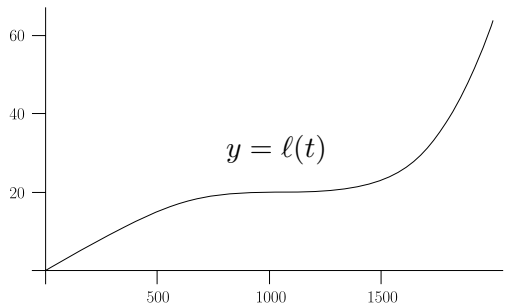


FIGURE 4.6.5

Computational Note: Theoretically, the least squares solutions of $\mathbf{Ax} = \mathbf{b}$ are exactly the solutions of the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, but forming and solving the normal equations to compute least squares solutions with floating-point arithmetic is not recommended. As pointed out in Example 4.5.1 on p. 214, any sensitivities to small perturbations that are present in the underlying problem are magnified by forming the normal equations. In other words, if the underlying problem is somewhat ill-conditioned, then the system of normal equations will be ill-conditioned to an even greater extent. Numerically stable techniques that avoid the normal equations are presented in Example 5.5.3 on p. 313 and Example 5.7.3 on p. 346.

Epilogue

While viewing a region in the Taurus constellation on January 1, 1801, Giuseppe Piazzi, an astronomer and director of the Palermo observatory, observed a small "star" that he had never seen before. As Piazzi and others continued to watch this new "star"—which was really an asteroid—they noticed that it was in fact moving, and they concluded that a new "planet" had been discovered. However, their new "planet" completely disappeared in the autumn of 1801. Well-known astronomers of the time joined the search to relocate the lost "planet," but all efforts were in vain.

In September of 1801 Carl F. Gauss decided to take up the challenge of finding this lost "planet." Gauss allowed for the possibility of an elliptical orbit rather than constraining it to be circular—which was an assumption of the others—and he proceeded to develop the method of least squares. By December the task was completed, and Gauss informed the scientific community not only where the lost "planet" was located, but he also predicted its position at future times. They looked, and it was exactly where Gauss had predicted it would be! The asteroid was named *Ceres*, and Gauss's contribution was recognized by naming another minor asteroid *Gaussia.*

This extraordinary feat of locating a tiny and distant heavenly body from apparently insufficient data astounded the scientific community. Furthermore, Gauss refused to reveal his methods, and there were those who even accused him of sorcery. These events led directly to Gauss's fame throughout the entire European community, and they helped to establish his reputation as a mathematical and scientific genius of the highest order.

Gauss waited until 1809, when he published his *Theoria Motus Corporum Coelestium In Sectionibus Conicis Solem Ambientium,* to systematically develop the theory of least squares and his methods of orbit calculation. This was in keeping with Gauss's philosophy to publish nothing but well-polished work of lasting significance. When criticized for not revealing more motivational aspects in his writings, Gauss remarked that architects of great cathedrals do not obscure the beauty of their work by leaving the scaffolds in place after the construction has been completed. Gauss's theory of least squares approximation has indeed proven to be a great mathematical cathedral of lasting beauty and significance.

Exercises for section 4.6

4.6.1. Hooke's law says that the displacement y of an ideal spring is proportional to the force x that is applied—i.e., $y = kx$ for some constant k. Consider a spring in which k is unknown. Various masses are attached, and the resulting displacements shown in Figure 4.6.6 are observed. Using these observations, determine the least squares estimate for k.

x (lb)	y (in)
5	11.1
7	15.4
8	17.5
10	22.0
12	26.3

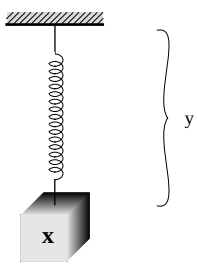


Figure 4.6.6

4.6.2. Show that the slope of the line that passes through the origin in $\Re^2$ and comes closest in the least squares sense to passing through the points $\{(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)\}$ is given by $m = \sum_i x_i y_i / \sum_i x_i^2$.

4.6.3. A small company has been in business for three years and has recorded annual profits (in thousands of dollars) as follows.

Year	1	2	3
Sales	7	4	3

Assuming that there is a linear trend in the declining profits, predict the year and the month in which the company begins to lose money.

4.6.4. An economist hypothesizes that the change (in dollars) in the price of a loaf of bread is primarily a linear combination of the change in the price of a bushel of wheat and the change in the minimum wage. That is, if B is the change in bread prices, W is the change in wheat prices, and M is the change in the minimum wage, then $B = \alpha W + \beta M$. Suppose that for three consecutive years the change in bread prices, wheat prices, and the minimum wage are as shown below.

	Year 1	Year 2	Year 3
B	+\$1	+\$1	+\$1
W	+\$1	+\$2	0\$
M	+\$1	0\$	−\$1

Use the theory of least squares to estimate the change in the price of bread in Year 4 if wheat prices and the minimum wage each fall by \$1.

4.6.5. Suppose that a researcher hypothesizes that the weight loss of a pint of ice cream during storage is primarily a linear function of time. That is,

$$y = \alpha_0 + \alpha_1 t + \varepsilon,$$

where $y =$ the weight loss in grams, $t =$ the storage time in weeks, and ε is a random error function whose mean value is 0. Suppose that an experiment is conducted, and the following data is obtained.

Time (t)	1	2	3	4	5	6	7	8
Loss (y)	.15	.21	.30	.41	.49	.59	.72	.83

(a) Determine the least squares estimates for the parameters α_0 and α_1.

(b) Predict the mean weight loss for a pint of ice cream that is stored for 20 weeks.

4.6.6. After studying a certain type of cancer, a researcher hypothesizes that in the short run the number (y) of malignant cells in a particular tissue grows exponentially with time (t). That is, $y = \alpha_0 e^{\alpha_1 t}$. Determine least squares estimates for the parameters α_0 and α_1 from the researcher's observed data given below.

t (days)	1	2	3	4	5
y (cells)	16	27	45	74	122

Hint: What common transformation converts an exponential function into a linear function?

4.6.7. Using least squares techniques, fit the following data

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
y	2	7	9	12	13	14	14	13	10	8	4

with a line $y = \alpha_0 + \alpha_1 x$ and then fit the data with a quadratic $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$. Determine which of these two curves best fits the data by computing the sum of the squares of the errors in each case.

4.6.8. Consider the time (T) it takes for a runner to complete a marathon (26 miles and 385 yards). Many factors such as height, weight, age, previous training, etc. can influence an athlete's performance, but experience has shown that the following three factors are particularly important:

$$x_1 = \text{Ponderal index} = \frac{\text{height (in.)}}{[\text{weight (lbs.)}]^{\frac{1}{3}}},$$

$$x_2 = \text{Miles run the previous 8 weeks},$$

$$x_3 = \text{Age (years)}.$$

A linear model hypothesizes that the time T (in minutes) is given by $T = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon$, where ε is a random function accounting for all other factors and whose mean value is assumed to be zero. On the basis of the five observations given below, estimate the expected marathon time for a 43-year-old runner of height 74 in., weight 180 lbs., who has run 450 miles during the previous eight weeks.

T	x_1	x_2	x_3
181	13.1	619	23
193	13.5	803	42
212	13.8	207	31
221	13.1	409	38
248	12.5	482	45

What is your personal predicted mean marathon time?

4.6.9. For $\mathbf{A} \in \Re^{m\times n}$ and $\mathbf{b} \in \Re^m$, prove that $\mathbf{x}_2$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}_2$ is part of a solution to the larger system

$$\begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \tag{4.6.5}$$

Note: It is not uncommon to encounter least squares problems in which $\mathbf{A}$ is extremely large but very sparse (mostly zero entries). For these situations, the system (4.6.5) will usually contain significantly fewer nonzero entries than the system of normal equations, thereby helping to overcome the memory requirements that plague these problems. Using (4.6.5) also eliminates the undesirable need to explicitly form the product $\mathbf{A}^T\mathbf{A}$—recall from Example 4.5.1 that forming $\mathbf{A}^T\mathbf{A}$ can cause loss of significant information.

4.6.10. In many least squares applications, the underlying data matrix $\mathbf{A}_{m\times n}$ does not have independent columns—i.e., $rank\,(\mathbf{A}) < n$—so the corresponding system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ will fail to have a unique solution. This means that in an associated linear estimation problem of the form

$$y = \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_n t_n + \varepsilon$$

there will be infinitely many least squares estimates for the parameters α_i, and hence there will be infinitely many estimates for the mean value of y at any given point $(t_1, t_2, \ldots, t_n)$—which is clearly an undesirable situation. In order to remedy this problem, we restrict ourselves to making estimates only at those points $(t_1, t_2, \ldots, t_n)$ that are in the row space of $\mathbf{A}$. If

$$\mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} \in R\left(\mathbf{A}^T\right), \quad \text{and if} \quad \mathbf{x} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix}$$

is any least squares solution (i.e., $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$), prove that the estimate defined by

$$\hat{y} = \mathbf{t}^T\mathbf{x} = \sum_{i=1}^{n} t_i \hat{\alpha}_i$$

is unique in the sense that $\hat{y}$ is independent of which least squares solution $\mathbf{x}$ is used.

4.7 LINEAR TRANSFORMATIONS

The connection between linear functions and matrices is at the heart of our subject. As explained on p. 93, matrix algebra grew out of Cayley's observation that the composition of two linear functions can be represented by the multiplication of two matrices. It's now time to look deeper into such matters and to formalize the connections between matrices, vector spaces, and linear functions defined on vector spaces. This is the point at which linear algebra, as the study of linear functions on vector spaces, begins in earnest.

Linear Transformations

Let $\mathcal{U}$ and $\mathcal{V}$ be vector spaces over a field $\mathcal{F}$ ($\Re$ or $\mathcal{C}$ for us).

- A ***linear transformation*** from $\mathcal{U}$ into $\mathcal{V}$ is defined to be a linear function $\mathbf{T}$ mapping $\mathcal{U}$ into $\mathcal{V}$. That is,

$$\mathbf{T}(\mathbf{x}+\mathbf{y}) = \mathbf{T}(\mathbf{x}) + \mathbf{T}(\mathbf{y}) \quad \text{and} \quad \mathbf{T}(\alpha\mathbf{x}) = \alpha\mathbf{T}(\mathbf{x}) \tag{4.7.1}$$

  or, equivalently,

$$\mathbf{T}(\alpha\mathbf{x}+\mathbf{y}) = \alpha\mathbf{T}(\mathbf{x}) + \mathbf{T}(\mathbf{y}) \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathcal{U},\ \alpha \in \mathcal{F}. \tag{4.7.2}$$

- A ***linear operator*** on $\mathcal{U}$ is defined to be a linear transformation from $\mathcal{U}$ into itself—i.e., a linear function mapping $\mathcal{U}$ back into $\mathcal{U}$.

Example 4.7.1

- The function $\mathbf{0}(\mathbf{x}) = \mathbf{0}$ that maps all vectors in a space $\mathcal{U}$ to the zero vector in another space $\mathcal{V}$ is a linear transformation from $\mathcal{U}$ into $\mathcal{V}$, and, not surprisingly, it is called the ***zero transformation.***
- The function $\mathbf{I}(\mathbf{x}) = \mathbf{x}$ that maps every vector from a space $\mathcal{U}$ back to itself is a linear operator on $\mathcal{U}$. $\mathbf{I}$ is called the ***identity operator*** on $\mathcal{U}$.
- For $\mathbf{A} \in \Re^{m\times n}$ and $\mathbf{x} \in \Re^{n\times 1}$, the function $\mathbf{T}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is a linear transformation from $\Re^n$ into $\Re^m$ because matrix multiplication satisfies $\mathbf{A}(\alpha\mathbf{x}+\mathbf{y}) = \alpha\mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y}$. $\mathbf{T}$ is a linear operator on $\Re^n$ if $\mathbf{A}$ is $n \times n$.
- If $\mathcal{W}$ is the vector space of all functions from $\Re$ to $\Re$, and if $\mathcal{V}$ is the space of all differentiable functions from $\Re$ to $\Re$, then the mapping $\mathbf{D}(f) = df/dx$ is a linear transformation from $\mathcal{V}$ into $\mathcal{W}$ because

$$\frac{d(\alpha f + g)}{dx} = \alpha\frac{df}{dx} + \frac{dg}{dx}.$$

- If $\mathcal{V}$ is the space of all continuous functions from $\Re$ into $\Re$, then the mapping defined by $\mathbf{T}(f) = \int_0^x f(t)dt$ is a linear operator on $\mathcal{V}$ because

$$\int_0^x [\alpha f(t) + g(t)]\,dt = \alpha\int_0^x f(t)dt + \int_0^x g(t)dt.$$

- The ***rotator*** $\mathbf{Q}$ that rotates vectors $\mathbf{u}$ in $\Re^2$ counterclockwise through an angle θ, as shown in Figure 4.7.1, is a linear operator on $\Re^2$ because the "action" of $\mathbf{Q}$ on $\mathbf{u}$ can be described by matrix multiplication in the sense that the coordinates of the rotated vector $\mathbf{Q}(\mathbf{u})$ are given by

$$\mathbf{Q}(\mathbf{u}) = \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

- The ***projector*** $\mathbf{P}$ that maps each point $\mathbf{v} = (x, y, z) \in \Re^3$ to its orthogonal projection $(x, y, 0)$ in the xy-plane, as depicted in Figure 4.7.2, is a linear operator on $\Re^3$ because if $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, then

$$\mathbf{P}(\alpha\mathbf{u} + \mathbf{v}) = (\alpha u_1 + v_1, \alpha u_2 + v_2, 0) = \alpha(u_1, u_2, 0) + (v_1, v_2, 0) = \alpha\mathbf{P}(\mathbf{u}) + \mathbf{P}(\mathbf{v}).$$

- The ***reflector*** $\mathbf{R}$ that maps each vector $\mathbf{v} = (x, y, z) \in \Re^3$ to its reflection $\mathbf{R}(\mathbf{v}) = (x, y, -z)$ about the xy-plane, as shown in Figure 4.7.3, is a linear operator on $\Re^3$.

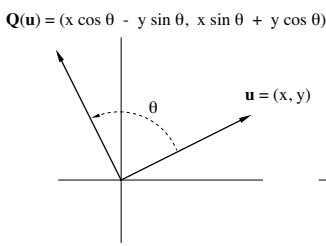


Figure 4.7.1

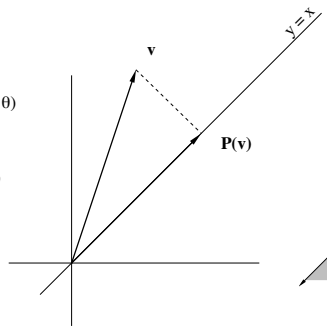


Figure 4.7.2

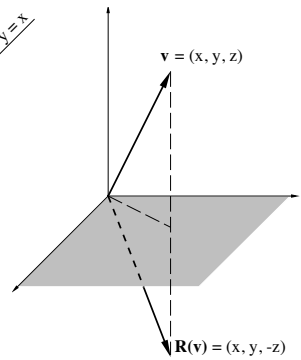


Figure 4.7.3

- Just as the rotator $\mathbf{Q}$ is represented by a matrix $[\mathbf{Q}] = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, the projector $\mathbf{P}$ and the reflector $\mathbf{R}$ can be represented by matrices

$$[\mathbf{P}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad [\mathbf{R}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

in the sense that the "action" of $\mathbf{P}$ and $\mathbf{R}$ on $\mathbf{v} = (x, y, z)$ can be accomplished with matrix multiplication using $[\mathbf{P}]$ and $[\mathbf{R}]$ by writing

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ -z \end{pmatrix}.$$

It would be wrong to infer from Example 4.7.1 that all linear transformations can be represented by matrices (of finite size). For example, the differential and integral operators do not have matrix representations because they are defined on infinite-dimensional spaces. But linear transformations on *finite*-dimensional spaces will always have matrix representations. To see why, the concept of "coordinates" in higher dimensions must first be understood.

Recall that if $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is a basis for a vector space $\mathcal{U}$, then each $\mathbf{v} \in \mathcal{U}$ can be written as $\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n$. The α_i's in this expansion are uniquely determined by $\mathbf{v}$ because if $\mathbf{v} = \sum_i \alpha_i\mathbf{u}_i = \sum_i \beta_i\mathbf{u}_i$, then $\mathbf{0} = \sum_i(\alpha_i - \beta_i)\mathbf{u}_i$, and this implies $\alpha_i - \beta_i = 0$ (i.e., $\alpha_i = \beta_i$) for each i because $\mathcal{B}$ is an independent set.

Coordinates of a Vector

Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be a basis for a vector space $\mathcal{U}$, and let $\mathbf{v} \in \mathcal{U}$. The coefficients α_i in the expansion $\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n$ are called the ***coordinates of*** $\mathbf{v}$ ***with respect to*** $\mathcal{B}$, and, from now on, $[\mathbf{v}]_{\mathcal{B}}$ will denote the column vector

$$[\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

Caution! Order is important. If $\mathcal{B}'$ is a permutation of $\mathcal{B}$, then $[\mathbf{v}]_{\mathcal{B}'}$ is the corresponding permutation of $[\mathbf{v}]_{\mathcal{B}}$.

From now on, $\mathcal{S} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ will denote the standard basis of unit vectors (in natural order) for $\Re^n$ (or $\mathcal{C}^n$). If no other basis is explicitly mentioned, then the standard basis is assumed. For example, if no basis is mentioned, and if we write

$$\mathbf{v} = \begin{pmatrix} 8 \\ 7 \\ 4 \end{pmatrix},$$

then it is understood that this is the representation with respect to $\mathcal{S}$ in the sense that $\mathbf{v} = [\mathbf{v}]_{\mathcal{S}} = 8\mathbf{e}_1 + 7\mathbf{e}_2 + 4\mathbf{e}_3$. The ***standard coordinates*** of a vector are its coordinates with respect to $\mathcal{S}$. So, 8, 7, and 4 are the standard coordinates of $\mathbf{v}$ in the above example.

Example 4.7.2

Problem: If $\mathbf{v}$ is a vector in $\Re^3$ whose standard coordinates are

$$\mathbf{v} = \begin{pmatrix} 8 \\ 7 \\ 4 \end{pmatrix},$$

determine the coordinates of $\mathbf{v}$ with respect to the basis

$$\mathcal{B} = \left\{ \mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\}.$$

Solution: The object is to find the three unknowns $\alpha_1, \alpha_2,$ and α_3 such that $\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3 = \mathbf{v}$. This is simply a 3×3 system of linear equations

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 4 \end{pmatrix} \implies [\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} 9 \\ 2 \\ -3 \end{pmatrix}.$$

The general rule for making a change of coordinates is given on p. 252.

Linear transformations possess coordinates in the same way vectors do because linear transformations from $\mathcal{U}$ to $\mathcal{V}$ also form a vector space.

Space of Linear Transformations

- For each pair of vector spaces $\mathcal{U}$ and $\mathcal{V}$ over $\mathcal{F}$, the set $\mathcal{L}(\mathcal{U}, \mathcal{V})$ of all linear transformations from $\mathcal{U}$ to $\mathcal{V}$ is a vector space over $\mathcal{F}$.
- Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ and $\mathcal{B}' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ be bases for $\mathcal{U}$ and $\mathcal{V}$, respectively, and let $\mathbf{B}_{ji}$ be the linear transformation from $\mathcal{U}$ into $\mathcal{V}$ defined by $\mathbf{B}_{ji}(\mathbf{u}) = \xi_j \mathbf{v}_i$, where $(\xi_1, \xi_2, \ldots, \xi_n)^T = [\mathbf{u}]_{\mathcal{B}}$. That is, pick off the jth coordinate of $\mathbf{u}$, and attach it to $\mathbf{v}_i$.
  - ▷ $\mathcal{B}_{\mathcal{L}} = \{\mathbf{B}_{ji}\}_{j=1\ldots n}^{i=1\ldots m}$ is a basis for $\mathcal{L}(\mathcal{U}, \mathcal{V})$.
  - ▷ $\dim \mathcal{L}(\mathcal{U}, \mathcal{V}) = (\dim \mathcal{U})(\dim \mathcal{V})$.

Proof. $\mathcal{L}(\mathcal{U}, \mathcal{V})$ is a vector space because the defining properties on p. 160 are satisfied—details are omitted. Prove $\mathcal{B}_{\mathcal{L}}$ is a basis by demonstrating that it is a linearly independent spanning set for $\mathcal{L}(\mathcal{U}, \mathcal{V})$. To establish linear independence, suppose $\sum_{j,i} \eta_{ji}\mathbf{B}_{ji} = \mathbf{0}$ for scalars η_{ji}, and observe that for each $\mathbf{u}_k \in \mathcal{B}$,

$$\mathbf{B}_{ji}(\mathbf{u}_k) = \begin{cases} \mathbf{v}_i & \text{if } j = k \\ \mathbf{0} & \text{if } j \neq k \end{cases} \implies \mathbf{0} = \Big(\sum_{j,i} \eta_{ji}\mathbf{B}_{ji}\Big)(\mathbf{u}_k) = \sum_{j,i} \eta_{ji}\mathbf{B}_{ji}(\mathbf{u}_k) = \sum_{i=1}^{m} \eta_{ki}\mathbf{v}_i.$$

For each k, the independence of $\mathcal{B}'$ implies that $\eta_{ki} = 0$ for each i, and thus $\mathcal{B}_{\mathcal{L}}$ is linearly independent. To see that $\mathcal{B}_{\mathcal{L}}$ spans $\mathcal{L}(\mathcal{U}, \mathcal{V})$, let $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$,

and determine the action of $\mathbf{T}$ on any $\mathbf{u} \in \mathcal{U}$ by using $\mathbf{u} = \sum_{j=1}^{n} \xi_j \mathbf{u}_j$ and $\mathbf{T}(\mathbf{u}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{v}_i$ to write

$$\begin{aligned}\mathbf{T}(\mathbf{u}) &= \mathbf{T}\left(\sum_{j=1}^{n} \xi_j \mathbf{u}_j\right) = \sum_{j=1}^{n} \xi_j \mathbf{T}(\mathbf{u}_j) = \sum_{j=1}^{n} \xi_j \sum_{i=1}^{m} \alpha_{ij}\mathbf{v}_i \\ &= \sum_{i,j} \alpha_{ij}\xi_j \mathbf{v}_i = \sum_{i,j} \alpha_{ij}\mathbf{B}_{ji}(\mathbf{u}).\end{aligned} \tag{4.7.3}$$

This holds for all $\mathbf{u} \in \mathcal{U}$, so $\mathbf{T} = \sum_{i,j} \alpha_{ij}\mathbf{B}_{ji}$, and thus $\mathcal{B}_{\mathcal{L}}$ spans $\mathcal{L}(\mathcal{U}, \mathcal{V})$. ∎

It now makes sense to talk about the *coordinates* of $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ with respect to the basis $\mathcal{B}_{\mathcal{L}}$. In fact, the rule for determining these coordinates is contained in the proof above, where it was demonstrated that $\mathbf{T} = \sum_{i,j} \alpha_{ij}\mathbf{B}_{ji}$ in which the coordinates α_{ij} are precisely the scalars in

$$\mathbf{T}(\mathbf{u}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{v}_i \text{ or, equivalently, } [\mathbf{T}(\mathbf{u}_j)]_{\mathcal{B}'} = \begin{pmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{mj} \end{pmatrix}, \quad j = 1, 2, \ldots, n.$$

This suggests that rather than listing all coordinates α_{ij} in a single column containing mn entries (as we did with coordinate vectors), it's more logical to arrange the α_{ij}'s as an $m \times n$ matrix in which the jth column contains the coordinates of $\mathbf{T}(\mathbf{u}_j)$ with respect to $\mathcal{B}'$. These ideas are summarized below.

Coordinate Matrix Representations

Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ and $\mathcal{B}' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$ be bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. The ***coordinate matrix*** of $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ with respect to the pair $(\mathcal{B}, \mathcal{B}')$ is defined to be the $m \times n$ matrix

$$[\mathbf{T}]_{\mathcal{B}\mathcal{B}'} = \Big([\mathbf{T}(\mathbf{u}_1)]_{\mathcal{B}'} \;\Big|\; [\mathbf{T}(\mathbf{u}_2)]_{\mathcal{B}'} \;\Big|\; \cdots \;\Big|\; [\mathbf{T}(\mathbf{u}_n)]_{\mathcal{B}'} \Big). \tag{4.7.4}$$

In other words, if $\mathbf{T}(\mathbf{u}_j) = \alpha_{1j}\mathbf{v}_1 + \alpha_{2j}\mathbf{v}_2 + \cdots + \alpha_{mj}\mathbf{v}_m$, then

$$[\mathbf{T}(\mathbf{u}_j)]_{\mathcal{B}'} = \begin{pmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{mj} \end{pmatrix} \text{ and } [\mathbf{T}]_{\mathcal{B}\mathcal{B}'} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix}. \tag{4.7.5}$$

When $\mathbf{T}$ is a linear operator on $\mathcal{U}$, and when there is only one basis involved, $[\mathbf{T}]_{\mathcal{B}}$ is used in place of $[\mathbf{T}]_{\mathcal{B}\mathcal{B}}$ to denote the (necessarily square) coordinate matrix of $\mathbf{T}$ with respect to $\mathcal{B}$.

Example 4.7.3

Problem: If $\mathbf{P}$ is the projector defined in Example 4.7.1 that maps each point $\mathbf{v} = (x, y, z) \in \Re^3$ to its orthogonal projection $\mathbf{P}(\mathbf{v}) = (x, y, 0)$ in the xy-plane, determine the coordinate matrix $[\mathbf{P}]_{\mathcal{B}}$ with respect to the basis

$$\mathcal{B} = \left\{ \mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\}.$$

Solution: According to (4.7.4), the jth column in $[\mathbf{P}]_{\mathcal{B}}$ is $[\mathbf{P}(\mathbf{u}_j)]_{\mathcal{B}}$. Therefore,

$$\mathbf{P}(\mathbf{u}_1) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = 1\mathbf{u}_1 + 1\mathbf{u}_2 - 1\mathbf{u}_3 \implies [\mathbf{P}(\mathbf{u}_1)]_{\mathcal{B}} = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix},$$

$$\mathbf{P}(\mathbf{u}_2) = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} = 0\mathbf{u}_1 + 3\mathbf{u}_2 - 2\mathbf{u}_3 \implies [\mathbf{P}(\mathbf{u}_2)]_{\mathcal{B}} = \begin{pmatrix} 0 \\ 3 \\ -2 \end{pmatrix},$$

$$\mathbf{P}(\mathbf{u}_3) = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} = 0\mathbf{u}_1 + 3\mathbf{u}_2 - 2\mathbf{u}_3 \implies [\mathbf{P}(\mathbf{u}_3)]_{\mathcal{B}} = \begin{pmatrix} 0 \\ 3 \\ -2 \end{pmatrix},$$

so that $[\mathbf{P}]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & 3 \\ -1 & -2 & -2 \end{pmatrix}$.

Example 4.7.4

Problem: Consider the same problem given in Example 4.7.3, but use different bases—say,

$$\mathcal{B} = \left\{ \mathbf{u}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$$

and

$$\mathcal{B}' = \left\{ \mathbf{v}_1 = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\}.$$

For the projector defined by $\mathbf{P}(x, y, z) = (x, y, 0)$, determine $[\mathbf{P}]_{\mathcal{B}\mathcal{B}'}$.

Solution: Determine the coordinates of each $\mathbf{P}(\mathbf{u}_j)$ with respect to $\mathcal{B}'$, as

shown below:

$$\mathbf{P}(\mathbf{u}_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = -1\mathbf{v_1} + 0\mathbf{v_2} + 0\mathbf{v_3} \implies [\mathbf{P}(\mathbf{u}_1)]_{\mathcal{B}'} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix},$$

$$\mathbf{P}(\mathbf{u}_2) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = -1\mathbf{v_1} + 1\mathbf{v_2} + 0\mathbf{v_3} \implies [\mathbf{P}(\mathbf{u}_2)]_{\mathcal{B}'} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix},$$

$$\mathbf{P}(\mathbf{u}_3) = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = -1\mathbf{v_1} + 1\mathbf{v_2} + 0\mathbf{v_3} \implies [\mathbf{P}(\mathbf{u}_3)]_{\mathcal{B}'} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}.$$

Therefore, according to (4.7.4), $[\mathbf{P}]_{\mathcal{B}\mathcal{B}'} = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$.

At the heart of linear algebra is the realization that the theory of finite-dimensional linear transformations is essentially the same as the theory of matrices. This is due primarily to the fundamental fact that the action of a linear transformation $\mathbf{T}$ on a vector $\mathbf{u}$ is precisely matrix multiplication between the coordinates of $\mathbf{T}$ and the coordinates of $\mathbf{u}$.

Action as Matrix Multiplication

Let $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$, and let $\mathcal{B}$ and $\mathcal{B}'$ be bases for $\mathcal{U}$ and $\mathcal{V}$, respectively. For each $\mathbf{u} \in \mathcal{U}$, the action of $\mathbf{T}$ on $\mathbf{u}$ is given by matrix multiplication between their coordinates in the sense that

$$[\mathbf{T}(\mathbf{u})]_{\mathcal{B}'} = [\mathbf{T}]_{\mathcal{B}\mathcal{B}'}[\mathbf{u}]_{\mathcal{B}}. \tag{4.7.6}$$

Proof. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ and $\mathcal{B}' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$. If $\mathbf{u} = \sum_{j=1}^{n} \xi_j \mathbf{u}_j$ and $\mathbf{T}(\mathbf{u}_j) = \sum_{i=1}^{m} \alpha_{ij}\mathbf{v}_i$, then

$$[\mathbf{u}]_{\mathcal{B}} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} \quad \text{and} \quad [\mathbf{T}]_{\mathcal{B}\mathcal{B}'} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix},$$

so, according to (4.7.3),

$$\mathbf{T}(\mathbf{u}) = \sum_{i,j} \alpha_{ij}\xi_j \mathbf{v}_i = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij}\xi_j \right) \mathbf{v}_i.$$

In other words, the coordinates of $\mathbf{T}(\mathbf{u})$ with respect to $\mathcal{B}'$ are the terms $\sum_{j=1}^{n} \alpha_{ij}\xi_j$ for $i = 1, 2, \ldots, m$, and therefore

$$[\mathbf{T}(\mathbf{u})]_{\mathcal{B}'} = \begin{pmatrix} \sum_j \alpha_{1j}\xi_j \\ \sum_j \alpha_{2j}\xi_j \\ \vdots \\ \sum_j \alpha_{mj}\xi_j \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} = [\mathbf{T}]_{\mathcal{B}\mathcal{B}'}[\mathbf{u}]_{\mathcal{B}}. \quad \blacksquare$$

Example 4.7.5

Problem: Show how the action of the operator $\mathbf{D}\big(p(t)\big) = dp/dt$ on the space $\mathcal{P}_3$ of polynomials of degree three or less is given by matrix multiplication.

Solution: The coordinate matrix of $\mathbf{D}$ with respect to the basis $\mathcal{B} = \{1, t, t^2, t^3\}$ is

$$[\mathbf{D}]_{\mathcal{B}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

If $\mathbf{p} = p(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$, then $\mathbf{D}(\mathbf{p}) = \alpha_1 + 2\alpha_2 t + 3\alpha_3 t^2$ so that

$$[\mathbf{p}]_{\mathcal{B}} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \quad \text{and} \quad [\mathbf{D}(\mathbf{p})]_{\mathcal{B}} = \begin{pmatrix} \alpha_1 \\ 2\alpha_2 \\ 3\alpha_3 \\ 0 \end{pmatrix}.$$

The action of $\mathbf{D}$ is accomplished by means of matrix multiplication because

$$[\mathbf{D}(\mathbf{p})]_{\mathcal{B}} = \begin{pmatrix} \alpha_1 \\ 2\alpha_2 \\ 3\alpha_3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = [\mathbf{D}]_{\mathcal{B}}[\mathbf{p}]_{\mathcal{B}}.$$

For $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ and $\mathbf{L} \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, the ***composition*** of $\mathbf{L}$ with $\mathbf{T}$ is defined to be the function $\mathbf{C} : \mathcal{U} \to \mathcal{W}$ such that $\mathbf{C}(\mathbf{x}) = \mathbf{L}\big(\mathbf{T}(\mathbf{x})\big)$, and this composition, denoted by $\mathbf{C} = \mathbf{LT}$, is also a linear transformation because

$$\begin{aligned} \mathbf{C}(\alpha\mathbf{x} + \mathbf{y}) &= \mathbf{L}\big(\mathbf{T}(\alpha\mathbf{x} + \mathbf{y})\big) = \mathbf{L}\big(\alpha\mathbf{T}(\mathbf{x}) + \mathbf{T}(\mathbf{y})\big) \\ &= \alpha\mathbf{L}\big(\mathbf{T}(\mathbf{x})\big) + \mathbf{L}\big(\mathbf{T}(\mathbf{y})\big) = \alpha\mathbf{C}(\mathbf{x}) + \mathbf{C}(\mathbf{y}). \end{aligned}$$

Consequently, if $\mathcal{B}$, $\mathcal{B}'$, and $\mathcal{B}''$ are bases for $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$, respectively, then $\mathbf{C}$ must have a coordinate matrix representation with respect to $(\mathcal{B}, \mathcal{B}'')$, so it's only natural to ask how $[\mathbf{C}]_{\mathcal{B}\mathcal{B}''}$ is related to $[\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}$ and $[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}$. Recall that the motivation behind the definition of matrix multiplication given on p. 93 was based on the need to represent the composition of two linear transformations, so it should be no surprise to discover that $[\mathbf{C}]_{\mathcal{B}\mathcal{B}''} = [\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}$. This, along with the other properties given below, makes it clear that studying linear transformations on finite-dimensional spaces amounts to studying matrix algebra.

Connections with Matrix Algebra

- If $\mathbf{T}, \mathbf{L} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$, and if $\mathcal{B}$ and $\mathcal{B}'$ are bases for $\mathcal{U}$ and $\mathcal{V}$, then
  - ▷ $[\alpha\mathbf{T}]_{\mathcal{B}\mathcal{B}'} = \alpha[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}$ for scalars α, (4.7.7)
  - ▷ $[\mathbf{T}+\mathbf{L}]_{\mathcal{B}\mathcal{B}'} = [\mathbf{T}]_{\mathcal{B}\mathcal{B}'} + [\mathbf{L}]_{\mathcal{B}\mathcal{B}'}$. (4.7.8)
- If $\mathbf{T} \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ and $\mathbf{L} \in \mathcal{L}(\mathcal{V}, \mathcal{W})$, and if $\mathcal{B}$, $\mathcal{B}'$, and $\mathcal{B}''$ are bases for $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$, respectively, then $\mathbf{LT} \in \mathcal{L}(\mathcal{U}, \mathcal{W})$, and
  - ▷ $[\mathbf{LT}]_{\mathcal{B}\mathcal{B}''} = [\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}$. (4.7.9)
- If $\mathbf{T} \in \mathcal{L}(\mathcal{U},\mathcal{U})$ is invertible in the sense that $\mathbf{T}\mathbf{T}^{-1} = \mathbf{T}^{-1}\mathbf{T} = \mathbf{I}$ for some $\mathbf{T}^{-1} \in \mathcal{L}(\mathcal{U},\mathcal{U})$, then for every basis $\mathcal{B}$ of $\mathcal{U}$,
  - ▷ $[\mathbf{T}^{-1}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^{-1}$. (4.7.10)

Proof. The first three properties (4.7.7)–(4.7.9) follow directly from (4.7.6). For example, to prove (4.7.9), let $\mathbf{u}$ be any vector in $\mathcal{U}$, and write

$$[\mathbf{LT}]_{\mathcal{B}\mathcal{B}''}[\mathbf{u}]_{\mathcal{B}} = \big[\mathbf{LT}(\mathbf{u})\big]_{\mathcal{B}''} = \big[\mathbf{L}\big(\mathbf{T}(\mathbf{u})\big)\big]_{\mathcal{B}''} = [\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}\big[\mathbf{T}(\mathbf{u})\big]_{\mathcal{B}'} = [\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}[\mathbf{u}]_{\mathcal{B}}.$$

This is true for all $\mathbf{u} \in \mathcal{U}$, so $[\mathbf{LT}]_{\mathcal{B}\mathcal{B}''} = [\mathbf{L}]_{\mathcal{B}'\mathcal{B}''}[\mathbf{T}]_{\mathcal{B}\mathcal{B}'}$ (see Exercise 3.5.5). Proving (4.7.7) and (4.7.8) is similar—details are omitted. To prove (4.7.10), note that if $\dim\mathcal{U} = n$, then $[\mathbf{I}]_{\mathcal{B}} = \mathbf{I}_n$ for all bases $\mathcal{B}$, so property (4.7.9) implies $\mathbf{I}_n = [\mathbf{I}]_{\mathcal{B}} = [\mathbf{T}\mathbf{T}^{-1}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}[\mathbf{T}^{-1}]_{\mathcal{B}}$, and thus $[\mathbf{T}^{-1}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^{-1}$. ∎

Example 4.7.6

Problem: Form the composition $\mathbf{C} = \mathbf{LT}$ of the two linear transformations $\mathbf{T} : \Re^3 \to \Re^2$ and $\mathbf{L} : \Re^2 \to \Re^2$ defined by

$$\mathbf{T}(x, y, z) = (x + y,\ \ y - z) \quad \text{and} \quad \mathbf{L}(u, v) = (2u - v,\ \ u),$$

and then verify (4.7.9) and (4.7.10) using the standard bases $\mathcal{S}_2$ and $\mathcal{S}_3$ for $\Re^2$ and $\Re^3$, respectively.

Solution: The composition $\mathbf{C} : \Re^3 \to \Re^2$ is the linear transformation

$$\mathbf{C}(x, y, z) = \mathbf{L}\big(\mathbf{T}(x, y, z)\big) = \mathbf{L}(x + y,\ \ y - z) = (2x + y + z,\ \ x + y).$$

The coordinate matrix representations of $\mathbf{C}$, $\mathbf{L}$, and $\mathbf{T}$ are

$$[\mathbf{C}]_{\mathcal{S}_3\mathcal{S}_2} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \quad [\mathbf{L}]_{\mathcal{S}_2} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad [\mathbf{T}]_{\mathcal{S}_3\mathcal{S}_2} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Property (4.7.9) is verified because $[\mathbf{LT}]_{\mathcal{S}_3\mathcal{S}_2} = [\mathbf{C}]_{\mathcal{S}_3\mathcal{S}_2} = [\mathbf{L}]_{\mathcal{S}_2}[\mathbf{T}]_{\mathcal{S}_3\mathcal{S}_2}$. Find $\mathbf{L}^{-1}$ by looking for scalars β_{ij} in $\mathbf{L}^{-1}(u,v) = (\beta_{11}u + \beta_{12}v,\ \beta_{21}u + \beta_{22}v)$ such that $\mathbf{L}\mathbf{L}^{-1} = \mathbf{L}^{-1}\mathbf{L} = \mathbf{I}$ or, equivalently,

$$\mathbf{L}\big(\mathbf{L}^{-1}(u,v)\big) = \mathbf{L}^{-1}\big(\mathbf{L}(u,v)\big) = (u,v) \quad \text{for all } (u,v) \in \Re^2.$$

Computation reveals $\mathbf{L}^{-1}(u,v) = (v,\ 2v-u)$, and (4.7.10) is verified by noting

$$[\mathbf{L}^{-1}]_{\mathcal{S}_2} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}^{-1} = [\mathbf{L}]_{\mathcal{S}_2}^{-1}.$$

Exercises for section 4.7

4.7.1. Determine which of the following functions are linear operators on $\Re^2$.

(a) $\mathbf{T}(x,y) = (x, 1+y)$, (b) $\mathbf{T}(x,y) = (y,x)$,
(c) $\mathbf{T}(x,y) = (0, xy)$, (d) $\mathbf{T}(x,y) = (x^2, y^2)$,
(e) $\mathbf{T}(x,y) = (x, \sin y)$, (f) $\mathbf{T}(x,y) = (x+y, x-y)$.

4.7.2. For $\mathbf{A} \in \Re^{n\times n}$, determine which of the following functions are linear transformations.

(a) $\mathbf{T}(\mathbf{X}_{n\times n}) = \mathbf{A}\mathbf{X} - \mathbf{X}\mathbf{A}$, (b) $\mathbf{T}(\mathbf{x}_{n\times 1}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ for $\mathbf{b} \neq \mathbf{0}$,
(c) $\mathbf{T}(\mathbf{A}) = \mathbf{A}^T$, (d) $\mathbf{T}(\mathbf{X}_{n\times n}) = (\mathbf{X} + \mathbf{X}^T)/2$.

4.7.3. Explain why $\mathbf{T}(\mathbf{0}) = \mathbf{0}$ for every linear transformation $\mathbf{T}$.

4.7.4. Determine which of the following mappings are linear operators on $\mathcal{P}_n$, the vector space of polynomials of degree n or less.

(a) $\mathbf{T} = \xi_k\mathbf{D}^k + \xi_{k-1}\mathbf{D}^{k-1} + \cdots + \xi_1\mathbf{D} + \xi_0\mathbf{I}$, where $\mathbf{D}^k$ is the kth-order differentiation operator (i.e., $\mathbf{D}^k p(t) = d^k p/dt^k$).

(b) $\mathbf{T}\Big(p(t)\Big) = t^n p'(0) + t$.

4.7.5. Let $\mathbf{v}$ be a fixed vector in $\Re^{n\times 1}$ and let $\mathbf{T} : \Re^{n\times 1} \to \Re$ be the mapping defined by $\mathbf{T}(\mathbf{x}) = \mathbf{v}^{\mathbf{T}}\mathbf{x}$ (i.e., the standard inner product).

(a) Is $\mathbf{T}$ a linear operator?
(b) Is $\mathbf{T}$ a linear transformation?

4.7.6. For the operator $\mathbf{T} : \Re^2 \to \Re^2$ defined by $\mathbf{T}(x,y) = (x+y,\ -2x+4y)$, determine $[\mathbf{T}]_{\mathcal{B}}$, where $\mathcal{B}$ is the basis $\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$.

4.7.7. Let $\mathbf{T} : \Re^2 \to \Re^3$ be the linear transformation defined by $\mathbf{T}(x, y) = (x + 3y,\ 0,\ 2x - 4y)$.

(a) Determine $[\mathbf{T}]_{\mathcal{S}\mathcal{S}'}$, where $\mathcal{S}$ and $\mathcal{S}'$ are the standard bases for $\Re^2$ and $\Re^3$, respectively.

(b) Determine $[\mathbf{T}]_{\mathcal{S}\mathcal{S}''}$, where $\mathcal{S}''$ is the basis for $\Re^3$ obtained by permuting the standard basis according to $\mathcal{S}'' = \{\mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_1\}$.

4.7.8. Let $\mathbf{T}$ be the operator on $\Re^3$ defined by $\mathbf{T}(x, y, z) = (x-y,\ y-x,\ x-z)$ and consider the vector

$$\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} \quad \text{and the basis} \quad \mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}.$$

(a) Determine $[\mathbf{T}]_{\mathcal{B}}$ and $[\mathbf{v}]_{\mathcal{B}}$.

(b) Compute $[\mathbf{T}(\mathbf{v})]_{\mathcal{B}}$, and then verify that $[\mathbf{T}]_{\mathcal{B}}[\mathbf{v}]_{\mathcal{B}} = [\mathbf{T}(\mathbf{v})]_{\mathcal{B}}$.

4.7.9. For $\mathbf{A} \in \Re^{n\times n}$, let $\mathbf{T}$ be the linear operator on $\Re^{n\times 1}$ defined by $\mathbf{T}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. That is, $\mathbf{T}$ is the operator defined by matrix multiplication. With respect to the standard basis $\mathcal{S}$, show that $[\mathbf{T}]_{\mathcal{S}} = \mathbf{A}$.

4.7.10. If $\mathbf{T}$ is a linear operator on a space $\mathcal{V}$ with basis $\mathcal{B}$, explain why $[\mathbf{T}^k]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^k$ for all nonnegative integers k.

4.7.11. Let $\mathbf{P}$ be the projector that maps each point $\mathbf{v} \in \Re^2$ to its orthogonal projection on the line $y = x$ as depicted in Figure 4.7.4.

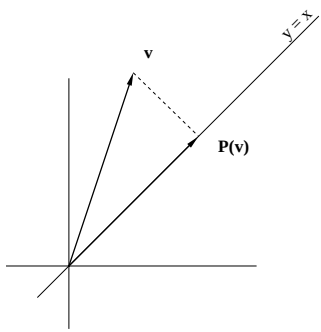


FIGURE 4.7.4

(a) Determine the coordinate matrix of $\mathbf{P}$ with respect to the standard basis.

(b) Determine the orthogonal projection of $\mathbf{v} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ onto the line $y = x$.

4.7.12. For the standard basis $\mathcal{S} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}$ of $\Re^{2\times 2}$, determine the matrix representation $[\mathbf{T}]_{\mathcal{S}}$ for each of the following linear operators on $\Re^{2\times 2}$, and then verify $[\mathbf{T}(\mathbf{U})]_{\mathcal{S}} = [\mathbf{T}]_{\mathcal{S}}[\mathbf{U}]_{\mathcal{S}}$ for $\mathbf{U} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

(a) $\mathbf{T}(\mathbf{X}_{2\times 2}) = \dfrac{\mathbf{X} + \mathbf{X}^T}{2}$.

(b) $\mathbf{T}(\mathbf{X}_{2\times 2}) = \mathbf{AX} - \mathbf{XA}$, where $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$.

4.7.13. For $\mathcal{P}_2$ and $\mathcal{P}_3$ (the spaces of polynomials of degrees less than or equal to two and three, respectively), let $\mathbf{S} : \mathcal{P}_2 \to \mathcal{P}_3$ be the linear transformation defined by $\mathbf{S}(p) = \int_0^t p(x)dx$. Determine $[\mathbf{S}]_{\mathcal{B}\mathcal{B}'}$, where $\mathcal{B} = \{1, t, t^2\}$ and $\mathcal{B}' = \{1, t, t^2, t^3\}$.

4.7.14. Let $\mathbf{Q}$ be the linear operator on $\Re^2$ that rotates each point counterclockwise through an angle θ, and let $\mathbf{R}$ be the linear operator on $\Re^2$ that reflects each point about the x-axis.

(a) Determine the matrix of the composition $[\mathbf{RQ}]_{\mathcal{S}}$ relative to the standard basis $\mathcal{S}$.

(b) Relative to the standard basis, determine the matrix of the linear operator that rotates each point in $\Re^2$ counterclockwise through an angle 2θ.

4.7.15. Let $\mathbf{P} : \mathcal{U} \to \mathcal{V}$ and $\mathbf{Q} : \mathcal{U} \to \mathcal{V}$ be two linear transformations, and let $\mathcal{B}$ and $\mathcal{B}'$ be arbitrary bases for $\mathcal{U}$ and $\mathcal{V}$, respectively.

(a) Provide the details to explain why $[\mathbf{P}+\mathbf{Q}]_{\mathcal{B}\mathcal{B}'} = [\mathbf{P}]_{\mathcal{B}\mathcal{B}'} + [\mathbf{Q}]_{\mathcal{B}\mathcal{B}'}$.

(b) Provide the details to explain why $[\alpha\mathbf{P}]_{\mathcal{B}\mathcal{B}'} = \alpha[\mathbf{P}]_{\mathcal{B}\mathcal{B}'}$, where α is an arbitrary scalar.

4.7.16. Let $\mathbf{I}$ be the identity operator on an n-dimensional space $\mathcal{V}$.

(a) Explain why

$$[\mathbf{I}]_{\mathcal{B}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

regardless of the choice of basis $\mathcal{B}$.

(b) Let $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathcal{B}' = \{\mathbf{y}_i\}_{i=1}^n$ be two different bases for $\mathcal{V}$, and let $\mathbf{T}$ be the linear operator on $\mathcal{V}$ that maps vectors from $\mathcal{B}'$ to vectors in $\mathcal{B}$ according to the rule $\mathbf{T}(\mathbf{y}_i) = \mathbf{x}_i$ for $i = 1, 2, \ldots, n$. Explain why

$$[\mathbf{I}]_{\mathcal{B}\mathcal{B}'} = [\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}'} = \left([\mathbf{x}_1]_{\mathcal{B}'} \;\middle|\; [\mathbf{x}_2]_{\mathcal{B}'} \;\middle|\; \cdots \;\middle|\; [\mathbf{x}_n]_{\mathcal{B}'} \right).$$

(c) When $\mathcal{V} = \Re^3$, determine $[\mathbf{I}]_{\mathcal{B}\mathcal{B}'}$ for

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{B}' = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

4.7.17. Let $\mathbf{T} : \Re^3 \to \Re^3$ be the linear operator defined by

$$\mathbf{T}(x, y, z) = (2x - y, \ -x + 2y - z, \ z - y).$$

(a) Determine $\mathbf{T}^{-1}(x, y, z)$.
(b) Determine $[\mathbf{T}^{-1}]_{\mathcal{S}}$, where $\mathcal{S}$ is the standard basis for $\Re^3$.

4.7.18. Let $\mathbf{T}$ be a linear operator on an n-dimensional space $\mathcal{V}$. Show that the following statements are equivalent.
(1) $\mathbf{T^{-1}}$ exists.
(2) $\mathbf{T}$ is a one-to-one mapping (i.e., $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y}) \implies \mathbf{x} = \mathbf{y}$).
(3) $N(\mathbf{T}) = \{\mathbf{0}\}$.
(4) $\mathbf{T}$ is an onto mapping (i.e., for each $\mathbf{v} \in \mathcal{V}$, there is an $\mathbf{x} \in \mathcal{V}$ such that $\mathbf{T}(\mathbf{x}) = \mathbf{v}$).
Hint: Show that (1) $\implies$ (2) $\implies$ (3) $\implies$ (4) $\implies$ (2), and then show (2) and (4) $\implies$ (1).

4.7.19. Let $\mathcal{V}$ be an n-dimensional space with a basis $\mathcal{B} = \{\mathbf{u}_i\}_{i=1}^n$.
(a) Prove that a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\} \subseteq \mathcal{V}$ is linearly independent if and only if the set of coordinate vectors

$$\Big\{ [\mathbf{x}_1]_{\mathcal{B}}, \ [\mathbf{x}_2]_{\mathcal{B}}, \ \ldots, \ [\mathbf{x}_r]_{\mathcal{B}} \Big\} \subseteq \Re^{n \times 1}$$

is a linearly independent set.
(b) If $\mathbf{T}$ is a linear operator on $\mathcal{V}$, then the ***range*** of $\mathbf{T}$ is the set

$$R(\mathbf{T}) = \{\mathbf{T}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{V}\}.$$

Suppose that the basic columns of $[\mathbf{T}]_{\mathcal{B}}$ occur in positions $b_1, b_2, \ldots, b_r$. Explain why $\big\{\mathbf{T}(\mathbf{u}_{b_1}), \mathbf{T}(\mathbf{u}_{b_2}), \ldots, \mathbf{T}(\mathbf{u}_{b_r})\big\}$ is a basis for $R(\mathbf{T})$.

4.8 CHANGE OF BASIS AND SIMILARITY

By their nature, coordinate matrix representations are basis dependent. However, it's desirable to study linear transformations without reference to particular bases because some bases may force a coordinate matrix representation to exhibit special properties that are not present in the coordinate matrix relative to other bases. To divorce the study from the choice of bases it's necessary to somehow identify properties of coordinate matrices that are invariant among all bases—these are properties intrinsic to the transformation itself, and they are the ones on which to focus. The purpose of this section is to learn how to sort out these basis-independent properties.

The discussion is limited to a single finite-dimensional space $\mathcal{V}$ and to linear operators on $\mathcal{V}$. Begin by examining how the coordinates of $\mathbf{v} \in \mathcal{V}$ change as the basis for $\mathcal{V}$ changes. Consider two different bases

$$\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \quad \text{and} \quad \mathcal{B}' = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}.$$

It's convenient to regard $\mathcal{B}$ as an *old basis* for $\mathcal{V}$ and $\mathcal{B}'$ as a *new basis* for $\mathcal{V}$. Throughout this section $\mathbf{T}$ will denote the linear operator such that

$$\mathbf{T}(\mathbf{y}_i) = \mathbf{x}_i \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{4.8.1}$$

$\mathbf{T}$ is called the ***change of basis operator*** because it maps the new basis vectors in $\mathcal{B}'$ to the old basis vectors in $\mathcal{B}$. Notice that $[\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}'} = [\mathbf{I}]_{\mathcal{B}\mathcal{B}'}$. To see this, observe that

$$\mathbf{x}_i = \sum_{j=1}^{n} \alpha_j \mathbf{y}_j \quad \Longrightarrow \quad \mathbf{T}(\mathbf{x}_i) = \sum_{j=1}^{n} \alpha_j \mathbf{T}(\mathbf{y}_j) = \sum_{j=1}^{n} \alpha_j \mathbf{x}_j,$$

which means $[\mathbf{x}_i]_{\mathcal{B}'} = [\mathbf{T}(\mathbf{x}_i)]_{\mathcal{B}}$, so, according to (4.7.4),

$$[\mathbf{T}]_{\mathcal{B}} = \Big([\mathbf{T}(\mathbf{x}_1)]_{\mathcal{B}} \; [\mathbf{T}(\mathbf{x}_2)]_{\mathcal{B}} \; \cdots [\mathbf{T}(\mathbf{x}_n)]_{\mathcal{B}} \Big) = \Big([\mathbf{x}_1]_{\mathcal{B}'} \; [\mathbf{x}_2]_{\mathcal{B}'} \cdots [\mathbf{x}_n]_{\mathcal{B}'} \Big) = [\mathbf{T}]_{\mathcal{B}'}.$$

The fact that $[\mathbf{I}]_{\mathcal{B}\mathcal{B}'} = [\mathbf{T}]_{\mathcal{B}}$ follows because $[\mathbf{I}(\mathbf{x}_i)]_{\mathcal{B}'} = [\mathbf{x}_i]_{\mathcal{B}'}$. The matrix

$$\mathbf{P} = [\mathbf{I}]_{\mathcal{B}\mathcal{B}'} = [\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}'} \tag{4.8.2}$$

will hereafter be referred to as a ***change of basis matrix.*** **Caution!** $[\mathbf{I}]_{\mathcal{B}\mathcal{B}'}$ is not necessarily the identity matrix—see Exercise 4.7.16—and $[\mathbf{I}]_{\mathcal{B}\mathcal{B}'} \neq [\mathbf{I}]_{\mathcal{B}'\mathcal{B}}$.

We are now in a position to see how the coordinates of a vector change as the basis for the underlying space changes.

Changing Vector Coordinates

Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\mathcal{B}' = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ be bases for $\mathcal{V}$, and let $\mathbf{T}$ and $\mathbf{P}$ be the associated change of basis operator and change of basis matrix, respectively—i.e., $\mathbf{T}(\mathbf{y}_i) = \mathbf{x}_i$, for each i, and

$$\mathbf{P} = [\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}'} = [\mathbf{I}]_{\mathcal{B}\mathcal{B}'} = \Big([\mathbf{x}_1]_{\mathcal{B}'} \,\Big|\, [\mathbf{x}_2]_{\mathcal{B}'} \,\Big|\, \cdots \,\Big|\, [\mathbf{x}_n]_{\mathcal{B}'} \Big). \tag{4.8.3}$$

- $[\mathbf{v}]_{\mathcal{B}'} = \mathbf{P}[\mathbf{v}]_{\mathcal{B}}$ for all $\mathbf{v} \in \mathcal{V}$. (4.8.4)
- $\mathbf{P}$ is nonsingular.
- No other matrix can be used in place of $\mathbf{P}$ in (4.8.4).

Proof. Use (4.7.6) to write $[\mathbf{v}]_{\mathcal{B}'} = [\mathbf{I}(\mathbf{v})]_{\mathcal{B}'} = [\mathbf{I}]_{\mathcal{B}\mathcal{B}'}[\mathbf{v}]_{\mathcal{B}} = \mathbf{P}[\mathbf{v}]_{\mathcal{B}}$, which is (4.8.4). $\mathbf{P}$ is nonsingular because $\mathbf{T}$ is invertible (in fact, $\mathbf{T}^{-1}(\mathbf{x}_i) = \mathbf{y}_i$), and because (4.7.10) insures $[\mathbf{T}^{-1}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^{-1} = \mathbf{P}^{-1}$. $\mathbf{P}$ is unique because if $\mathbf{W}$ is another matrix satisfying (4.8.4) for all $\mathbf{v} \in \mathcal{V}$, then $(\mathbf{P} - \mathbf{W})[\mathbf{v}]_{\mathcal{B}} = \mathbf{0}$ for all $\mathbf{v}$. Taking $\mathbf{v} = \mathbf{x}_i$ yields $(\mathbf{P} - \mathbf{W})\mathbf{e}_i = \mathbf{0}$ for each i, so $\mathbf{P} - \mathbf{W} = \mathbf{0}$. ∎

If we think of $\mathcal{B}$ as the *old* basis and $\mathcal{B}'$ as the *new* basis, then the change of basis operator $\mathbf{T}$ acts as

$$\mathbf{T}(\text{new basis}) = \text{old basis},$$

while the change of basis matrix $\mathbf{P}$ acts as

$$\text{new coordinates} = \mathbf{P}(\text{old coordinates}).$$

For this reason, $\mathbf{T}$ should be referred to as the change of basis operator *from* $\mathcal{B}'$ *to* $\mathcal{B}$, while $\mathbf{P}$ is called the change of basis matrix *from* $\mathcal{B}$ *to* $\mathcal{B}'$.

Example 4.8.1

Problem: For the space $\mathcal{P}_2$ of polynomials of degree 2 or less, determine the change of basis matrix $\mathbf{P}$ from $\mathcal{B}$ to $\mathcal{B}'$, where

$$\mathcal{B} = \{1, t, t^2\} \quad \text{and} \quad \mathcal{B}' = \{1,\ 1+t,\ 1+t+t^2\},$$

and then find the coordinates of $q(t) = 3 + 2t + 4t^2$ relative to $\mathcal{B}'$.

Solution: According to (4.8.3), the change of basis matrix from $\mathcal{B}$ to $\mathcal{B}'$ is

$$\mathbf{P} = \Big([\mathbf{x}_1]_{\mathcal{B}'} \,\Big|\, [\mathbf{x}_2]_{\mathcal{B}'} \,\Big|\, [\mathbf{x}_3]_{\mathcal{B}'} \Big).$$

In this case, $\mathbf{x}_1 = 1$, $\mathbf{x}_2 = t$, and $\mathbf{x}_3 = t^2$, and $\mathbf{y}_1 = 1$, $\mathbf{y}_2 = 1 + t$, and $\mathbf{y}_3 = 1 + t + t^2$, so the coordinates $[\mathbf{x}_i]_{\mathcal{B}'}$ are computed as follows:

$$\begin{aligned} 1 &= 1(1) + 0(1+t) + 0(1+t+t^2) = 1\mathbf{y}_1 + 0\mathbf{y}_2 + 0\mathbf{y}_3, \\ t &= -1(1) + 1(1+t) + 0(1+t+t^2) = -1\mathbf{y}_1 + 1\mathbf{y}_2 + 0\mathbf{y}_3, \\ t^2 &= 0(1) - 1(1+t) + 1(1+t+t^2) = 0\mathbf{y}_1 - 1\mathbf{y}_2 + 1\mathbf{y}_3. \end{aligned}$$

Therefore,

$$\mathbf{P} = \left([\mathbf{x}_1]_{\mathcal{B}'} \;\middle|\; [\mathbf{x}_2]_{\mathcal{B}'} \;\middle|\; [\mathbf{x}_3]_{\mathcal{B}'} \right) = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

and the coordinates of $\mathbf{q} = q(t) = 3 + 2t + 4t^2$ with respect to $\mathcal{B}'$ are

$$[\mathbf{q}]_{\mathcal{B}'} = \mathbf{P}[\mathbf{q}]_{\mathcal{B}} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 4 \end{pmatrix}.$$

To independently check that these coordinates are correct, simply verify that

$$q(t) = 1(1) - 2(1+t) + 4(1+t+t^2).$$

It's now rather easy to describe how the coordinate matrix of a linear operator changes as the underlying basis changes.

Changing Matrix Coordinates

Let $\mathbf{A}$ be a linear operator on $\mathcal{V}$, and let $\mathcal{B}$ and $\mathcal{B}'$ be two bases for $\mathcal{V}$. The coordinate matrices $[\mathbf{A}]_{\mathcal{B}}$ and $[\mathbf{A}]_{\mathcal{B}'}$ are related as follows.

$$[\mathbf{A}]_{\mathcal{B}} = \mathbf{P}^{-1}[\mathbf{A}]_{\mathcal{B}'}\mathbf{P}, \quad \text{where} \quad \mathbf{P} = [\mathbf{I}]_{\mathcal{B}\mathcal{B}'} \tag{4.8.5}$$

is the change of basis matrix from $\mathcal{B}$ to $\mathcal{B}'$. Equivalently,

$$[\mathbf{A}]_{\mathcal{B}'} = \mathbf{Q}^{-1}[\mathbf{A}]_{\mathcal{B}}\mathbf{Q}, \quad \text{where} \quad \mathbf{Q} = [\mathbf{I}]_{\mathcal{B}'\mathcal{B}} = \mathbf{P}^{-1} \tag{4.8.6}$$

is the change of basis matrix from $\mathcal{B}'$ to $\mathcal{B}$.

Proof. Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\mathcal{B}' = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, and observe that for each j, (4.7.6) can be used to write

$$\Big[\mathbf{A}(\mathbf{x}_j)\Big]_{\mathcal{B}'} = [\mathbf{A}]_{\mathcal{B}'}\,[\mathbf{x}_j]_{\mathcal{B}'} = [\mathbf{A}]_{\mathcal{B}'}\mathbf{P}_{*j} = \Big[[\mathbf{A}]_{\mathcal{B}'}\mathbf{P}\Big]_{*j}.$$

Now use the change of coordinates rule (4.8.4) together with the fact that $\big[\mathbf{A}(\mathbf{x}_j)\big]_{\mathcal{B}} = \big[[\mathbf{A}]_{\mathcal{B}}\big]_{*j}$ (see (4.7.4)) to write

$$\Big[\mathbf{A}(\mathbf{x}_j)\Big]_{\mathcal{B}'} = \mathbf{P}\Big[\mathbf{A}(\mathbf{x}_j)\Big]_{\mathcal{B}} = \mathbf{P}\Big[[\mathbf{A}]_{\mathcal{B}}\Big]_{*j} = \Big[\mathbf{P}[\mathbf{A}]_{\mathcal{B}}\Big]_{*j}.$$

Consequently, $\big[[\mathbf{A}]_{\mathcal{B}'}\mathbf{P}\big]_{*j} = \big[\mathbf{P}[\mathbf{A}]_{\mathcal{B}}\big]_{*j}$ for each j, so $[\mathbf{A}]_{\mathcal{B}'}\mathbf{P} = \mathbf{P}[\mathbf{A}]_{\mathcal{B}}$. Since the change of basis matrix $\mathbf{P}$ is nonsingular, it follows that $[\mathbf{A}]_{\mathcal{B}} = \mathbf{P}^{-1}[\mathbf{A}]_{\mathcal{B}'}\mathbf{P}$, and (4.8.5) is proven. Setting $\mathbf{Q} = \mathbf{P}^{-1}$ in (4.8.5) yields $[\mathbf{A}]_{\mathcal{B}'} = \mathbf{Q}^{-1}[\mathbf{A}]_{\mathcal{B}}\mathbf{Q}$. The matrix $\mathbf{Q} = \mathbf{P}^{-1}$ is the change of basis matrix from $\mathcal{B}'$ to $\mathcal{B}$ because if $\mathbf{T}$ is the change of basis operator from $\mathcal{B}'$ to $\mathcal{B}$ (i.e., $\mathbf{T}(\mathbf{y}_i) = \mathbf{x}_i$), then $\mathbf{T}^{-1}$ is the change of basis operator from $\mathcal{B}$ to $\mathcal{B}'$ (i.e., $\mathbf{T}^{-1}(\mathbf{x}_i) = \mathbf{y}_i$), and according to (4.8.3), the change of basis matrix from $\mathcal{B}'$ to $\mathcal{B}$ is

$$[\mathbf{I}]_{\mathcal{B}'\mathcal{B}} = \Big(\ [\mathbf{y}_1]_{\mathcal{B}}\ \Big|\ [\mathbf{y}_2]_{\mathcal{B}}\ \Big|\ \cdots\ \Big|\ [\mathbf{y}_n]_{\mathcal{B}}\Big) = [\mathbf{T}^{-1}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^{-1} = \mathbf{P}^{-1} = \mathbf{Q}. \quad \blacksquare$$

Example 4.8.2

Problem: Consider the linear operator $\mathbf{A}(x, y) = (y,\ -2x + 3y)$ on $\Re^2$ along with the two bases

$$\mathcal{S} = \left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\} \quad \text{and} \quad \mathcal{S}' = \left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\right\}.$$

First compute the coordinate matrix $[\mathbf{A}]_{\mathcal{S}}$ as well as the change of basis matrix $\mathbf{Q}$ from $\mathcal{S}'$ to $\mathcal{S}$, and then use these two matrices to determine $[\mathbf{A}]_{\mathcal{S}'}$.

Solution: The matrix of $\mathbf{A}$ relative to $\mathcal{S}$ is obtained by computing

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

so that $[\mathbf{A}]_{\mathcal{S}} = \big(\ [\mathbf{A}(\mathbf{e}_1)]_{\mathcal{S}}\ \big|\ [\mathbf{A}(\mathbf{e}_2)]_{\mathcal{S}}\ \big) = \begin{pmatrix}0 & 1\\ -2 & 3\end{pmatrix}$. According to (4.8.6), the change of basis matrix from $\mathcal{S}'$ to $\mathcal{S}$ is

$$\mathbf{Q} = \Big(\ [\mathbf{y}_1]_{\mathcal{S}}\ \Big|\ [\mathbf{y}_2]_{\mathcal{S}}\ \Big) = \begin{pmatrix}1 & 1\\ 1 & 2\end{pmatrix},$$

and the matrix of $\mathbf{A}$ with respect to $\mathcal{S}'$ is

$$[\mathbf{A}]_{\mathcal{S}'} = \mathbf{Q}^{-1}[\mathbf{A}]_{\mathcal{S}}\mathbf{Q} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -2 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Notice that $[\mathbf{A}]_{\mathcal{S}'}$ is a diagonal matrix, whereas $[\mathbf{A}]_{\mathcal{S}}$ is not. This shows that the standard basis is not always the best choice for providing a simple matrix representation. Finding a basis so that the associated coordinate matrix is as simple as possible is one of the fundamental issues of matrix theory. Given an operator $\mathbf{A}$, the solution to the general problem of determining a basis $\mathcal{B}$ so that $[\mathbf{A}]_{\mathcal{B}}$ is diagonal is summarized on p. 520.

Example 4.8.3

Problem: Consider a matrix $\mathbf{M}_{n\times n}$ to be a linear operator on $\Re^n$ by defining $\mathbf{M}(\mathbf{v}) = \mathbf{M}\mathbf{v}$ (matrix–vector multiplication). If $\mathcal{S}$ is the standard basis for $\Re^n$, and if $\mathcal{S}' = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ is any other basis, describe $[\mathbf{M}]_{\mathcal{S}}$ and $[\mathbf{M}]_{\mathcal{S}'}$.

Solution: The jth column in $[\mathbf{M}]_{\mathcal{S}}$ is $[\mathbf{M}\mathbf{e}_j]_{\mathcal{S}} = [\mathbf{M}_{*j}]_{\mathcal{S}} = \mathbf{M}_{*j}$, and hence $[\mathbf{M}]_{\mathcal{S}} = \mathbf{M}$. That is, the coordinate matrix of $\mathbf{M}$ with respect to $\mathcal{S}$ is $\mathbf{M}$ itself. To find $[\mathbf{M}]_{\mathcal{S}'}$, use (4.8.6) to write $[\mathbf{M}]_{\mathcal{S}'} = \mathbf{Q}^{-1}[\mathbf{M}]_{\mathcal{S}}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}$, where

$$\mathbf{Q} = [\mathbf{I}]_{\mathcal{S}'\mathcal{S}} = \Big([\mathbf{q}_1]_{\mathcal{S}} \,\big|\, [\mathbf{q}_2]_{\mathcal{S}} \,\big|\, \cdots \,\big|\, [\mathbf{q}_n]_{\mathcal{S}} \Big) = \Big(\mathbf{q}_1 \,\big|\, \mathbf{q}_2 \,\big|\, \cdots \,\big|\, \mathbf{q}_n\Big).$$

Conclusion: The matrices $\mathbf{M}$ and $\mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}$ represent the same linear operator (namely, $\mathbf{M}$), but with respect to two different bases (namely, $\mathcal{S}$ and $\mathcal{S}'$). So, when considering properties of $\mathbf{M}$ (as a linear operator), it's legitimate to replace $\mathbf{M}$ by $\mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}$. Whenever the structure of $\mathbf{M}$ obscures its operator properties, look for a basis $\mathcal{S}' = \{\mathbf{Q}_{*1}, \mathbf{Q}_{*2}, \ldots, \mathbf{Q}_{*n}\}$ (or, equivalently, a nonsingular matrix $\mathbf{Q}$) such that $\mathbf{Q}^{-1}\mathbf{M}\mathbf{Q}$ has a simpler structure. This is an important theme throughout linear algebra and matrix theory.

For a linear operator $\mathbf{A}$, the special relationships between $[\mathbf{A}]_{\mathcal{B}}$ and $[\mathbf{A}]_{\mathcal{B}'}$ that are given in (4.8.5) and (4.8.6) motivate the following definitions.

Similarity

- Matrices $\mathbf{B}_{n\times n}$ and $\mathbf{C}_{n\times n}$ are said to be ***similar matrices*** whenever there exists a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}$. We write $\mathbf{B} \simeq \mathbf{C}$ to denote that $\mathbf{B}$ and $\mathbf{C}$ are similar.
- The linear operator $f : \Re^{n\times n} \to \Re^{n\times n}$ defined by $f(\mathbf{C}) = \mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}$ is called a ***similarity transformation.***

Equations (4.8.5) and (4.8.6) say that any two coordinate matrices of a given linear operator must be similar. But must any two similar matrices be coordinate matrices of the same linear operator? Yes, and here's why. Suppose $\mathbf{C} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$, and let $\mathbf{A}(\mathbf{v}) = \mathbf{B}\mathbf{v}$ be the linear operator defined by matrix–vector multiplication. If $\mathcal{S}$ is the standard basis, then it's straightforward to see that $[\mathbf{A}]_{\mathcal{S}} = \mathbf{B}$ (Exercise 4.7.9). If $\mathcal{B}' = \{\mathbf{Q}_{*1}, \mathbf{Q}_{*2}, \ldots, \mathbf{Q}_{*n}\}$ is the basis consisting of the columns of $\mathbf{Q}$, then (4.8.6) insures that $[\mathbf{A}]_{\mathcal{B}'} = [\mathbf{I}]^{-1}_{\mathcal{B}'\mathcal{S}}[\mathbf{A}]_{\mathcal{S}}[\mathbf{I}]_{\mathcal{B}'\mathcal{S}}$, where

$$[\mathbf{I}]_{\mathcal{B}'\mathcal{S}} = \Big([\mathbf{Q}_{*1}]_{\mathcal{S}} \,\Big|\, [\mathbf{Q}_{*2}]_{\mathcal{S}} \,\Big|\, \cdots \,\Big|\, [\mathbf{Q}_{*n}]_{\mathcal{S}} \Big) = \mathbf{Q}.$$

Therefore, $\mathbf{B} = [\mathbf{A}]_{\mathcal{S}}$ and $\mathbf{C} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q} = \mathbf{Q}^{-1}[\mathbf{A}]_{\mathcal{S}}\mathbf{Q} = [\mathbf{A}]_{\mathcal{B}'}$, so $\mathbf{B}$ and $\mathbf{C}$ are both coordinate matrix representations of $\mathbf{A}$. In other words, *similar matrices represent the same linear operator.*

As stated at the beginning of this section, the goal is to isolate and study coordinate-independent properties of linear operators. They are the ones determined by sorting out those properties of coordinate matrices that are basis independent. But, as (4.8.5) and (4.8.6) show, all coordinate matrices for a given linear operator must be similar, so the coordinate-independent properties are exactly the ones that are ***similarity invariant*** (invariant under similarity transformations). Naturally, determining and studying similarity invariants is an important part of linear algebra and matrix theory.

Example 4.8.4

Problem: The trace of a square matrix $\mathbf{C}$ was defined in Example 3.3.1 to be the sum of the diagonal entries

$$trace\,(\mathbf{C}) = \sum_i c_{ii}.$$

Show that trace is a similarity invariant, and explain why it makes sense to talk about the *trace of a linear operator* without regard to any particular basis. Then determine the trace of the linear operator on $\Re^2$ that is defined by

$$\mathbf{A}(x, y) = (y,\ -2x + 3y). \tag{4.8.7}$$

Solution: As demonstrated in Example 3.6.5, $trace\,(\mathbf{BC}) = trace\,(\mathbf{CB})$, whenever the products are defined, so

$$trace\,\big(\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}\big) = trace\,\big(\mathbf{C}\mathbf{Q}\mathbf{Q}^{-1}\big) = trace\,(\mathbf{C}),$$

and thus trace is a similarity invariant. This allows us to talk about the trace of a linear operator $\mathbf{A}$ without regard to any particular basis because $trace\,([\mathbf{A}]_{\mathcal{B}})$ is the same number regardless of the choice of $\mathcal{B}$. For example, two coordinate matrices of the operator $\mathbf{A}$ in (4.8.7) were computed in Example 4.8.2 to be

$$[\mathbf{A}]_{\mathcal{S}} = \begin{pmatrix} 0 & 1 \\ -2 & 3 \end{pmatrix} \quad \text{and} \quad [\mathbf{A}]_{\mathcal{S}'} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

and it's clear that $trace\,([\mathbf{A}]_{\mathcal{S}}) = trace\,([\mathbf{A}]_{\mathcal{S}'}) = 3$. Since $trace\,([\mathbf{A}]_{\mathcal{B}}) = 3$ for all $\mathcal{B}$, we can legitimately define $trace\,(\mathbf{A}) = 3$.

Exercises for section 4.8

4.8.1. Explain why rank is a similarity invariant.

4.8.2. Explain why similarity is transitive in the sense that $\mathbf{A} \simeq \mathbf{B}$ and $\mathbf{B} \simeq \mathbf{C}$ implies $\mathbf{A} \simeq \mathbf{C}$.

4.8.3. $\mathbf{A}(x, y, z) = (x + 2y - z,\ -y,\ x + 7z)$ is a linear operator on $\Re^3$.

(a) Determine $[\mathbf{A}]_{\mathcal{S}}$, where $\mathcal{S}$ is the standard basis.

(b) Determine $[\mathbf{A}]_{\mathcal{S}'}$ as well as the nonsingular matrix $\mathbf{Q}$ such that $[\mathbf{A}]_{\mathcal{S}'} = \mathbf{Q}^{-1}[\mathbf{A}]_{\mathcal{S}}\mathbf{Q}$ for $\mathcal{S}' = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$.

4.8.4. Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 4 \\ 0 & 1 & 5 \end{pmatrix}$ and $\mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\}$. Consider $\mathbf{A}$ as a linear operator on $\Re^{n \times 1}$ by means of matrix multiplication $\mathbf{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, and determine $[\mathbf{A}]_{\mathcal{B}}$.

4.8.5. Show that $\mathbf{C} = \begin{pmatrix} 4 & 6 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} -2 & -3 \\ 6 & 10 \end{pmatrix}$ are similar matrices, and find a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{C} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$. **Hint:** Consider $\mathbf{B}$ as a linear operator on $\Re^2$, and compute $[\mathbf{B}]_{\mathcal{S}}$ and $[\mathbf{B}]_{\mathcal{S}'}$, where $\mathcal{S}$ is the standard basis, and $\mathcal{S}' = \left\{ \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ 2 \end{pmatrix} \right\}$.

4.8.6. Let $\mathbf{T}$ be the linear operator $\mathbf{T}(x, y) = (-7x - 15y,\ 6x + 12y)$. Find a basis $\mathcal{B}$ such that $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$, and determine a matrix $\mathbf{Q}$ such that $[\mathbf{T}]_{\mathcal{B}} = \mathbf{Q}^{-1}[\mathbf{T}]_{\mathcal{S}}\mathbf{Q}$, where $\mathcal{S}$ is the standard basis.

4.8.7. By considering the rotator $\mathbf{P}(x, y) = (x\cos\theta - y\sin\theta,\ x\sin\theta + y\cos\theta)$ described in Example 4.7.1 and Figure 4.7.1, show that the matrices

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$$

are similar over the complex field. **Hint:** In case you have forgotten (or didn't know), $e^{i\theta} = \cos\theta + i\sin\theta$.

4.8.8. Let λ be a scalar such that $(\mathbf{C}-\lambda\mathbf{I})_{n\times n}$ is singular.

(a) If $\mathbf{B}\simeq\mathbf{C}$, prove that $(\mathbf{B}-\lambda\mathbf{I})$ is also singular.

(b) Prove that $(\mathbf{B}-\lambda_i\mathbf{I})$ is singular whenever $\mathbf{B}_{n\times n}$ is similar to

$$\mathbf{D}=\begin{pmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_n\end{pmatrix}.$$

4.8.9. If $\mathbf{A}\simeq\mathbf{B}$, show that $\mathbf{A}^k\simeq\mathbf{B}^k$ for all nonnegative integers k.

4.8.10. Suppose $\mathcal{B}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ and $\mathcal{B}'=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n\}$ are bases for an n-dimensional subspace $\mathcal{V}\subseteq\Re^{m\times 1}$, and let $\mathbf{X}_{m\times n}$ and $\mathbf{Y}_{m\times n}$ be the matrices whose columns are the vectors from $\mathcal{B}$ and $\mathcal{B}'$, respectively.

(a) Explain why $\mathbf{Y}^T\mathbf{Y}$ is nonsingular, and prove that the change of basis matrix from $\mathcal{B}$ to $\mathcal{B}'$ is $\mathbf{P}=\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{X}$.

(b) Describe $\mathbf{P}$ when $m=n$.

4.8.11. (a) $\mathbf{N}$ is *nilpotent of index* k when $\mathbf{N}^k=\mathbf{0}$ but $\mathbf{N}^{k-1}\neq\mathbf{0}$. If $\mathbf{N}$ is a nilpotent operator of index n on $\Re^n$, and if $\mathbf{N}^{n-1}(\mathbf{y})\neq\mathbf{0}$, show $\mathcal{B}=\left\{\mathbf{y},\mathbf{N}(\mathbf{y}),\mathbf{N}^2(\mathbf{y}),\ldots,\mathbf{N}^{n-1}(\mathbf{y})\right\}$ is a basis for $\Re^n$, and then demonstrate that

$$[\mathbf{N}]_{\mathcal{B}}=\mathbf{J}=\begin{pmatrix}0 & 0 & \cdots & 0 & 0\\ 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & 0\end{pmatrix}.$$

(b) If $\mathbf{A}$ and $\mathbf{B}$ are any two $n\times n$ nilpotent matrices of index n, explain why $\mathbf{A}\simeq\mathbf{B}$.

(c) Explain why all $n\times n$ nilpotent matrices of index n must have a zero trace and be of rank $n-1$.

4.8.12. $\mathbf{E}$ is *idempotent* when $\mathbf{E}^2=\mathbf{E}$. For an idempotent operator $\mathbf{E}$ on $\Re^n$, let $\mathcal{X}=\{\mathbf{x}_i\}_{i=1}^r$ and $\mathcal{Y}=\{\mathbf{y}_i\}_{i=1}^{n-r}$ be bases for $R(\mathbf{E})$ and $N(\mathbf{E})$, respectively.

(a) Prove that $\mathcal{B}=\mathcal{X}\cup\mathcal{Y}$ is a basis for $\Re^n$. **Hint:** Show $\mathbf{E}\mathbf{x}_i=\mathbf{x}_i$ and use this to deduce that $\mathcal{B}$ is linearly independent.

(b) Show that $[\mathbf{E}]_{\mathcal{B}}=\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$.

(c) Explain why two $n\times n$ idempotent matrices of the same rank must be similar.

(d) If $\mathbf{F}$ is an idempotent matrix, prove that $rank\,(\mathbf{F})=trace\,(\mathbf{F})$.

4.9 INVARIANT SUBSPACES

For a linear operator $\mathbf{T}$ on a vector space $\mathcal{V}$, and for $\mathcal{X} \subseteq \mathcal{V}$,

$$\mathbf{T}(\mathcal{X}) = \{\mathbf{T}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$$

is the set of all possible images of vectors from $\mathcal{X}$ under the transformation $\mathbf{T}$. Notice that $\mathbf{T}(\mathcal{V}) = R(\mathbf{T})$. When $\mathcal{X}$ is a subspace of $\mathcal{V}$, it follows that $\mathbf{T}(\mathcal{X})$ is also a subspace of $\mathcal{V}$, but $\mathbf{T}(\mathcal{X})$ is usually not related to $\mathcal{X}$. However, in some special cases it can happen that $\mathbf{T}(\mathcal{X}) \subseteq \mathcal{X}$, and such subspaces are the focus of this section.

Invariant Subspaces

- For a linear operator $\mathbf{T}$ on $\mathcal{V}$, a subspace $\mathcal{X} \subseteq \mathcal{V}$ is said to be an ***invariant subspace*** under $\mathbf{T}$ whenever $\mathbf{T}(\mathcal{X}) \subseteq \mathcal{X}$.
- In such a situation, $\mathbf{T}$ can be considered as a linear operator on $\mathcal{X}$ by forgetting about everything else in $\mathcal{V}$ and restricting $\mathbf{T}$ to act only on vectors from $\mathcal{X}$. Hereafter, this ***restricted operator*** will be denoted by $\mathbf{T}_{/\mathcal{X}}$.

Example 4.9.1

Problem: For

$$\mathbf{A} = \begin{pmatrix} 4 & 4 & 4 \\ -2 & -2 & -5 \\ 1 & 2 & 5 \end{pmatrix}, \qquad \mathbf{x}_1 = \begin{pmatrix} 2 \\ -1 \\ 0 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{x}_2 = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix},$$

show that the subspace $\mathcal{X}$ spanned by $\mathcal{B} = \{\mathbf{x}_1, \ \mathbf{x}_2\}$ is an invariant subspace under $\mathbf{A}$. Then describe the restriction $\mathbf{A}_{/\mathcal{X}}$ and determine the coordinate matrix of $\mathbf{A}_{/\mathcal{X}}$ relative to $\mathcal{B}$.

Solution: Observe that $\mathbf{A}\mathbf{x}_1 = 2\mathbf{x}_1 \in \mathcal{X}$ and $\mathbf{A}\mathbf{x}_2 = \mathbf{x}_1 + 2\mathbf{x}_2 \in \mathcal{X}$, so the image of any $\mathbf{x} = \alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in \mathcal{X}$ is back in $\mathcal{X}$ because

$$\mathbf{A}\mathbf{x} = \mathbf{A}(\alpha\mathbf{x}_1 + \beta\mathbf{x}_2) = \alpha\mathbf{A}\mathbf{x}_1 + \beta\mathbf{A}\mathbf{x}_2 = 2\alpha\mathbf{x}_1 + \beta(\mathbf{x}_1 + 2\mathbf{x}_2) = (2\alpha + \beta)\mathbf{x}_1 + 2\beta\mathbf{x}_2.$$

This equation completely describes the action of $\mathbf{A}$ restricted to $\mathcal{X}$, so

$$\mathbf{A}_{/\mathcal{X}}(\mathbf{x}) = (2\alpha + \beta)\mathbf{x}_1 + 2\beta\mathbf{x}_2 \quad \text{for each} \quad \mathbf{x} = \alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in \mathcal{X}.$$

Since $\mathbf{A}_{/\mathcal{X}}(\mathbf{x}_1) = 2\mathbf{x}_1$ and $\mathbf{A}_{/\mathcal{X}}(\mathbf{x}_2) = \mathbf{x}_1 + 2\mathbf{x}_2$, we have

$$\left[\mathbf{A}_{/\mathcal{X}}\right]_{\mathcal{B}} = \left(\left[\mathbf{A}_{/\mathcal{X}}(\mathbf{x}_1)\right]_{\mathcal{B}} \;\middle|\; \left[\mathbf{A}_{/\mathcal{X}}(\mathbf{x}_2)\right]_{\mathcal{B}} \right) = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}.$$

The invariant subspaces for a linear operator $\mathbf{T}$ are important because they produce simplified coordinate matrix representations of $\mathbf{T}$. To understand how this occurs, suppose $\mathcal{X}$ is an invariant subspace under $\mathbf{T}$, and let

$$\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$$

be a basis for $\mathcal{X}$ that is part of a basis

$$\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$$

for the entire space $\mathcal{V}$. To compute $[\mathbf{T}]_{\mathcal{B}}$, recall from the definition of coordinate matrices that

$$[\mathbf{T}]_{\mathcal{B}} = \Big([\mathbf{T}(\mathbf{x}_1)]_{\mathcal{B}} \,\Big|\, \cdots \,\Big|\, [\mathbf{T}(\mathbf{x}_r)]_{\mathcal{B}} \,\Big|\, [\mathbf{T}(\mathbf{y}_1)]_{\mathcal{B}} \,\Big|\, \cdots \,\Big|\, [\mathbf{T}(\mathbf{y}_q)]_{\mathcal{B}} \Big). \tag{4.9.1}$$

Because each $\mathbf{T}(\mathbf{x}_j)$ is contained in $\mathcal{X}$, only the first r vectors from $\mathcal{B}$ are needed to represent each $\mathbf{T}(\mathbf{x}_j)$, so, for $j = 1, 2, \ldots, r$,

$$\mathbf{T}(\mathbf{x}_j) = \sum_{i=1}^{r} \alpha_{ij}\mathbf{x}_i \quad \text{and} \quad [\mathbf{T}(\mathbf{x}_j)]_{\mathcal{B}} = \begin{pmatrix} \alpha_{1j} \\ \vdots \\ \alpha_{rj} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.9.2}$$

The space

$$\mathcal{Y} = span\,\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\} \tag{4.9.3}$$

may not be an invariant subspace for $\mathbf{T}$, so all the basis vectors in $\mathcal{B}$ may be needed to represent the $\mathbf{T}(\mathbf{y}_j)$'s. Consequently, for $j = 1, 2, \ldots, q$,

$$\mathbf{T}(\mathbf{y}_j) = \sum_{i=1}^{r} \beta_{ij}\mathbf{x}_i + \sum_{i=1}^{q} \gamma_{ij}\mathbf{y}_i \quad \text{and} \quad [\mathbf{T}(\mathbf{y}_j)]_{\mathcal{B}} = \begin{pmatrix} \beta_{1j} \\ \vdots \\ \beta_{rj} \\ \gamma_{1j} \\ \vdots \\ \gamma_{qj} \end{pmatrix}. \tag{4.9.4}$$

Using (4.9.2) and (4.9.4) in (4.9.1) produces the block-triangular matrix

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \alpha_{11} & \cdots & \alpha_{1r} & \beta_{11} & \cdots & \beta_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \cdots & \alpha_{rr} & \beta_{r1} & \cdots & \beta_{rq} \\ 0 & \cdots & 0 & \gamma_{11} & \cdots & \gamma_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \gamma_{q1} & \cdots & \gamma_{qq} \end{pmatrix}. \tag{4.9.5}$$

The equations $\mathbf{T}(\mathbf{x}_j) = \sum_{i=1}^{r} \alpha_{ij}\mathbf{x}_i$ in (4.9.2) mean that

$$\left[\mathbf{T}/_{\mathcal{X}}(\mathbf{x}_j)\right]_{\mathcal{B}_{\mathcal{X}}} = \begin{pmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{rj} \end{pmatrix}, \quad \text{so} \quad \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_{\mathcal{X}}} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1r} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rr} \end{pmatrix},$$

and thus the matrix in (4.9.5) can be written as

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_{\mathcal{X}}} & \mathbf{B}_{r\times q} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix}. \tag{4.9.6}$$

In other words, (4.9.6) says that the matrix representation for $\mathbf{T}$ can be made to be block triangular whenever a basis for an invariant subspace is available.

The more invariant subspaces we can find, the more tools we have to construct simplified matrix representations. For example, if the space $\mathcal{Y}$ in (4.9.3) is also an invariant subspace for $\mathbf{T}$, then $\mathbf{T}(\mathbf{y}_j) \in \mathcal{Y}$ for each $j = 1, 2, \ldots, q$, and only the $\mathbf{y}_i$'s are needed to represent $\mathbf{T}(\mathbf{y}_j)$ in (4.9.4). Consequently, the β_{ij} 's are all zero, and $[\mathbf{T}]_{\mathcal{B}}$ has the block-diagonal form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix} = \begin{pmatrix} \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_x} & \mathbf{0} \\ \mathbf{0} & \left[\mathbf{T}/_{\mathcal{Y}}\right]_{\mathcal{B}_y} \end{pmatrix}.$$

This notion easily generalizes in the sense that if $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}} \cup \cdots \cup \mathcal{B}_{\mathcal{Z}}$ is a basis for $\mathcal{V}$, where $\mathcal{B}_{\mathcal{X}}, \mathcal{B}_{\mathcal{Y}}, \ldots, \mathcal{B}_{\mathcal{Z}}$ are bases for invariant subspaces under $\mathbf{T}$ that have dimensions $r_1, r_2, \ldots, r_k$, respectively, then $[\mathbf{T}]_{\mathcal{B}}$ has the block-diagonal form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r_1\times r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{r_2\times r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{r_k\times r_k} \end{pmatrix},$$

where

$$\mathbf{A} = \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_x}, \quad \mathbf{B} = \left[\mathbf{T}/_{\mathcal{Y}}\right]_{\mathcal{B}_y}, \quad \ldots, \quad \mathbf{C} = \left[\mathbf{T}/_{\mathcal{Z}}\right]_{\mathcal{B}_z}.$$

The situations discussed above are also reversible in the sense that if the matrix representation of $\mathbf{T}$ has a block-triangular form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{B}_{r\times q} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix}$$

relative to some basis

$$\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q\},$$

then the r-dimensional subspace $\mathcal{U} = span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ spanned by the first r vectors in $\mathcal{B}$ must be an invariant subspace under $\mathbf{T}$. Furthermore, if the matrix representation of $\mathbf{T}$ has a block-diagonal form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix}$$

relative to $\mathcal{B}$, then both

$$\mathcal{U} = span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\} \quad \text{and} \quad \mathcal{W} = span\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q\}$$

must be invariant subspaces for $\mathbf{T}$. The details are left as exercises.

The general statement concerning invariant subspaces and coordinate matrix representations is given below.

Invariant Subspaces and Matrix Representations

Let $\mathbf{T}$ be a linear operator on an n-dimensional space $\mathcal{V}$, and let $\mathcal{X}, \mathcal{Y}, \ldots, \mathcal{Z}$ be subspaces of $\mathcal{V}$ with respective dimensions $r_1, r_2, \ldots, r_k$ and bases $\mathcal{B}_{\mathcal{X}}, \mathcal{B}_{\mathcal{Y}}, \ldots, \mathcal{B}_{\mathcal{Z}}$. Furthermore, suppose that $\sum_i r_i = n$ and $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}} \cup \cdots \cup \mathcal{B}_{\mathcal{Z}}$ is a basis for $\mathcal{V}$.

- The subspace $\mathcal{X}$ is an invariant subspace under $\mathbf{T}$ if and only if $[\mathbf{T}]_{\mathcal{B}}$ has the block-triangular form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r_1\times r_1} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \quad \text{in which case} \quad \mathbf{A} = \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_{\mathcal{X}}}. \tag{4.9.7}$$

- The subspaces $\mathcal{X}, \mathcal{Y}, \ldots, \mathcal{Z}$ are all invariant under $\mathbf{T}$ if and only if $[\mathbf{T}]_{\mathcal{B}}$ has the block-diagonal form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r_1\times r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{r_2\times r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{r_k\times r_k} \end{pmatrix}, \tag{4.9.8}$$

in which case

$$\mathbf{A} = \left[\mathbf{T}/_{\mathcal{X}}\right]_{\mathcal{B}_x}, \quad \mathbf{B} = \left[\mathbf{T}/_{\mathcal{Y}}\right]_{\mathcal{B}_y}, \quad \ldots, \quad \mathbf{C} = \left[\mathbf{T}/_{\mathcal{Z}}\right]_{\mathcal{B}_z}.$$

An important corollary concerns the special case in which the linear operator $\mathbf{T}$ is in fact an $n \times n$ matrix and $\mathbf{T}(\mathbf{v}) = \mathbf{T}\mathbf{v}$ is a matrix–vector multiplication.

Triangular and Diagonal Block Forms

When $\mathbf{T}$ is an $n \times n$ matrix, the following two statements are true.

- $\mathbf{Q}$ is a nonsingular matrix such that

$$\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{B}_{r\times q} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix} \tag{4.9.9}$$

  if and only if the first r columns in $\mathbf{Q}$ span an invariant subspace under $\mathbf{T}$.

- $\mathbf{Q}$ is a nonsingular matrix such that

$$\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q} = \begin{pmatrix} \mathbf{A}_{r_1\times r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{r_2\times r_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{r_k\times r_k} \end{pmatrix} \tag{4.9.10}$$

  if and only if $\mathbf{Q} = \left(\mathbf{Q}_1 \,\middle|\, \mathbf{Q}_2 \,\middle|\, \cdots \,\middle|\, \mathbf{Q}_k\right)$ in which $\mathbf{Q}_i$ is $n \times r_i$, and the columns of each $\mathbf{Q}_i$ span an invariant subspace under $\mathbf{T}$.

Proof. We know from Example 4.8.3 that if $\mathcal{B} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ is a basis for $\Re^n$, and if $\mathbf{Q} = \left(\mathbf{q}_1 \,\middle|\, \mathbf{q}_2 \,\middle|\, \cdots \,\middle|\, \mathbf{q}_n\right)$ is the matrix containing the vectors from $\mathcal{B}$ as its columns, then $[\mathbf{T}]_{\mathcal{B}} = \mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$. Statements (4.9.9) and (4.9.10) are now direct consequences of statements (4.9.7) and (4.9.8), respectively. ∎

Example 4.9.2

Problem: For

$$\mathbf{T} = \begin{pmatrix} -1 & -1 & -1 & -1 \\ 0 & -5 & -16 & -22 \\ 0 & 3 & 10 & 14 \\ 4 & 8 & 12 & 14 \end{pmatrix}, \quad \mathbf{q}_1 = \begin{pmatrix} 2 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{q}_2 = \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix},$$

verify that $\mathcal{X} = span\,\{\mathbf{q}_1, \mathbf{q}_2\}$ is an invariant subspace under $\mathbf{T}$, and then find a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$ has the block-triangular form

$$\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q} = \left(\begin{array}{cc|cc} * & * & * & * \\ * & * & * & * \\ \hline 0 & 0 & * & * \\ 0 & 0 & * & * \end{array}\right).$$

Solution: $\mathcal{X}$ is invariant because $\mathbf{T}\mathbf{q}_1 = \mathbf{q}_1 + 3\mathbf{q}_2$ and $\mathbf{T}\mathbf{q}_2 = 2\mathbf{q}_1 + 4\mathbf{q}_2$ insure that for all α and β, the images

$$\mathbf{T}(\alpha\mathbf{q}_1 + \beta\mathbf{q}_2) = (\alpha + 2\beta)\mathbf{q}_1 + (3\alpha + 4\beta)\mathbf{q}_2$$

lie in $\mathcal{X}$. The desired matrix $\mathbf{Q}$ is constructed by extending $\{\mathbf{q}_1, \mathbf{q}_2\}$ to a basis $\mathcal{B} = \{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4\}$ for $\Re^4$. If the extension technique described in Solution 2 of Example 4.4.5 is used, then

$$\mathbf{q}_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{q}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

and

$$\mathbf{Q} = \Big(\mathbf{q}_1 \mid \mathbf{q}_2 \mid \mathbf{q}_3 \mid \mathbf{q}_4\Big) = \begin{pmatrix} 2 & -1 & 1 & 0 \\ -1 & 2 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since the first two columns of $\mathbf{Q}$ span a space that is invariant under $\mathbf{T}$, it follows from (4.9.9) that $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$ must be in block-triangular form. This is easy to verify by computing

$$\mathbf{Q}^{-1} = \begin{pmatrix} 0 & -1 & -2 & 0 \\ 0 & 0 & -1 & 0 \\ 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{Q}^{-1}\mathbf{T}\mathbf{Q} = \left(\begin{array}{cc|cc} 1 & 2 & 0 & -6 \\ 3 & 4 & 0 & -14 \\ \hline 0 & 0 & -1 & -3 \\ 0 & 0 & 4 & 14 \end{array}\right).$$

In passing, notice that the upper-left-hand block is

$$\Big[\mathbf{T}_{/\mathcal{X}}\Big]_{\{\mathbf{q}_1, \mathbf{q}_2\}} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Example 4.9.3

Consider again the matrices of Example 4.9.2:

$$\mathbf{T} = \begin{pmatrix} -1 & -1 & -1 & -1 \\ 0 & -5 & -16 & -22 \\ 0 & 3 & 10 & 14 \\ 4 & 8 & 12 & 14 \end{pmatrix}, \quad \mathbf{q}_1 = \begin{pmatrix} 2 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{q}_2 = \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix}.$$

There are infinitely many extensions of $\{\mathbf{q}_1, \mathbf{q}_2\}$ to a basis $\mathcal{B} = \{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4\}$ for $\Re^4$—the extension used in Example 4.9.2 is only one possibility. Another extension is

$$\mathbf{q}_3 = \begin{pmatrix} 0 \\ -1 \\ 2 \\ -1 \end{pmatrix} \quad \text{and} \quad \mathbf{q}_4 = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}.$$

This extension might be preferred over that of Example 4.9.2 because the spaces $\mathcal{X} = span\{\mathbf{q}_1, \mathbf{q}_2\}$ and $\mathcal{Y} = span\{\mathbf{q}_3, \mathbf{q}_4\}$ are both invariant under $\mathbf{T}$, and therefore it follows from (4.9.10) that $\mathbf{Q}^{-1}\mathbf{TQ}$ is block diagonal. Indeed, it is not difficult to verify that

$$\mathbf{Q}^{-1}\mathbf{TQ} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} -1 & -1 & -1 & -1 \\ 0 & -5 & -16 & -22 \\ 0 & 3 & 10 & 14 \\ 4 & 8 & 12 & 14 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

$$= \left(\begin{array}{cc|cc} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ \hline 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{array}\right).$$

Notice that the diagonal blocks must be the matrices of the restrictions in the sense that

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \left[\mathbf{T}/_{\mathcal{X}}\right]_{\{\mathbf{q}_1, \mathbf{q}_2\}} \quad \text{and} \quad \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \left[\mathbf{T}/_{\mathcal{Y}}\right]_{\{\mathbf{q}_3, \mathbf{q}_4\}}.$$

Example 4.9.4

Problem: Find all subspaces of $\Re^2$ that are invariant under

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

Solution: The trivial subspace $\{\mathbf{0}\}$ is the only zero-dimensional invariant subspace, and the entire space $\Re^2$ is the only two-dimensional invariant subspace. The real problem is to find all one-dimensional invariant subspaces. If $\mathcal{M}$ is a one-dimensional subspace spanned by $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{A}(\mathcal{M}) \subseteq \mathcal{M}$, then

$$\mathbf{Ax} \in \mathcal{M} \implies \text{there is a scalar } \lambda \text{ such that } \mathbf{Ax} = \lambda\mathbf{x} \implies (\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}.$$

In other words, $\mathcal{M} \subseteq N(\mathbf{A} - \lambda\mathbf{I})$. Since $\dim \mathcal{M} = 1$, it must be the case that $N(\mathbf{A} - \lambda\mathbf{I}) \neq \{\mathbf{0}\}$, and consequently λ must be a scalar such that $(\mathbf{A} - \lambda\mathbf{I})$ is a singular matrix. Row operations produce

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} -\lambda & 1 \\ -2 & 3-\lambda \end{pmatrix} \longrightarrow \begin{pmatrix} -2 & 3-\lambda \\ -\lambda & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} -2 & 3-\lambda \\ 0 & 1+(\lambda^2-3\lambda)/2 \end{pmatrix},$$

and it is clear that $(\mathbf{A} - \lambda\mathbf{I})$ is singular if and only if $1 + (\lambda^2 - 3\lambda)/2 = 0$—i.e., if and only if λ is a root of

$$\lambda^2 - 3\lambda + 2 = 0.$$

Thus $\lambda = 1$ and $\lambda = 2$, and straightforward computation yields the two one-dimensional invariant subspaces

$$\mathcal{M}_1 = N(\mathbf{A} - \mathbf{I}) = span\left\{\begin{pmatrix}1\\1\end{pmatrix}\right\} \quad \text{and} \quad \mathcal{M}_2 = N(\mathbf{A} - \mathbf{2I}) = span\left\{\begin{pmatrix}1\\2\end{pmatrix}\right\}.$$

In passing, notice that $\mathcal{B} = \left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\right\}$ is a basis for $\Re^2$, and

$$[\mathbf{A}]_{\mathcal{B}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}, \quad \text{where} \quad \mathbf{Q} = \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}.$$

In general, scalars λ for which $(\mathbf{A} - \lambda\mathbf{I})$ is singular are called the ***eigenvalues*** of $\mathbf{A}$, and the nonzero vectors in $N(\mathbf{A} - \lambda\mathbf{I})$ are known as the associated ***eigenvectors*** for $\mathbf{A}$. As this example indicates, eigenvalues and eigenvectors are of fundamental importance in identifying invariant subspaces and reducing matrices by means of similarity transformations. Eigenvalues and eigenvectors are discussed at length in Chapter 7.

Exercises for section 4.9

4.9.1. Let $\mathbf{T}$ be an arbitrary linear operator on a vector space $\mathcal{V}$.
(a) Is the trivial subspace $\{\mathbf{0}\}$ invariant under $\mathbf{T}$?
(b) Is the entire space $\mathcal{V}$ invariant under $\mathbf{T}$?

4.9.2. Describe all of the subspaces that are invariant under the identity operator $\mathbf{I}$ on a space $\mathcal{V}$.

4.9.3. Let $\mathbf{T}$ be the linear operator on $\Re^4$ defined by

$$\mathbf{T}(x_1, x_2, x_3, x_4) = (x_1 + x_2 + 2x_3 - x_4, \quad x_2 + x_4, \quad 2x_3 - x_4, \quad x_3 + x_4),$$

and let $\mathcal{X} = span\{\mathbf{e}_1, \mathbf{e}_2\}$ be the subspace that is spanned by the first two unit vectors in $\Re^4$.
(a) Explain why $\mathcal{X}$ is invariant under $\mathbf{T}$.
(b) Determine $\left[\mathbf{T}/_{\mathcal{X}}\right]_{\{\mathbf{e}_1,\mathbf{e}_2\}}$.
(c) Describe the structure of $[\mathbf{T}]_{\mathcal{B}}$, where $\mathcal{B}$ is any basis obtained from an extension of $\{\mathbf{e}_1, \mathbf{e}_2\}$.

4.9.4. Let $\mathbf{T}$ and $\mathbf{Q}$ be the matrices

$$\mathbf{T} = \begin{pmatrix} -2 & -1 & -5 & -2 \\ -9 & 0 & -8 & -2 \\ 2 & 3 & 11 & 5 \\ 3 & -5 & -13 & -7 \end{pmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 1 & 1 & 3 & -4 \\ -2 & 0 & 1 & 0 \\ 3 & -1 & -4 & 3 \end{pmatrix}.$$

(a) Explain why the columns of $\mathbf{Q}$ are a basis for $\Re^4$.
(b) Verify that $\mathcal{X} = span\,\{\mathbf{Q}_{*1}, \mathbf{Q}_{*2}\}$ and $\mathcal{Y} = span\,\{\mathbf{Q}_{*3}, \mathbf{Q}_{*4}\}$ are each invariant subspaces under $\mathbf{T}$.
(c) Describe the structure of $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$ without doing any computation.
(d) Now compute the product $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$ to determine

$$\left[\mathbf{T}/_{\mathcal{X}}\right]_{\{\mathbf{Q}_{*1},\mathbf{Q}_{*2}\}} \quad \text{and} \quad \left[\mathbf{T}/_{\mathcal{Y}}\right]_{\{\mathbf{Q}_{*3},\mathbf{Q}_{*4}\}}.$$

4.9.5. Let $\mathbf{T}$ be a linear operator on a space $\mathcal{V}$, and suppose that

$$\mathcal{B} = \{\mathbf{u}_1, \ldots, \mathbf{u}_r, \ \mathbf{w}_1, \ldots, \mathbf{w}_q\}$$

is a basis for $\mathcal{V}$ such that $[\mathbf{T}]_{\mathcal{B}}$ has the block-diagonal form

$$[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix}.$$

Explain why $\mathcal{U} = span\,\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\mathcal{W} = span\,\{\mathbf{w}_1, \ldots, \mathbf{w}_q\}$ must each be invariant subspaces under $\mathbf{T}$.

4.9.6. If $\mathbf{T}_{n\times n}$ and $\mathbf{P}_{n\times n}$ are matrices such that

$$\mathbf{P}^{-1}\mathbf{T}\mathbf{P} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{q\times q} \end{pmatrix},$$

explain why

$$\mathcal{U} = span\,\{\mathbf{P}_{*1}, \ldots, \mathbf{P}_{*r}\} \quad \text{and} \quad \mathcal{W} = span\,\{\mathbf{P}_{*r+1}, \ldots, \mathbf{P}_{*n}\}$$

are each invariant subspaces under $\mathbf{T}$.

4.9.7. If $\mathbf{A}$ is an $n \times n$ matrix and λ is a scalar such that $(\mathbf{A} - \lambda\mathbf{I})$ is singular (i.e., λ is an eigenvalue), explain why the associated space of eigenvectors $N\,(\mathbf{A} - \lambda\mathbf{I})$ is an invariant subspace under $\mathbf{A}$.

4.9.8. Consider the matrix $\mathbf{A} = \begin{pmatrix} -9 & 4 \\ -24 & 11 \end{pmatrix}$.
(a) Determine the eigenvalues of $\mathbf{A}$.
(b) Identify all subspaces of $\Re^2$ that are invariant under $\mathbf{A}$.
(c) Find a nonsingular matrix $\mathbf{Q}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is a diagonal matrix.

CHAPTER 5

Norms, Inner Products, and Orthogonality

5.1 VECTOR NORMS

A significant portion of linear algebra is in fact geometric in nature because much of the subject grew out of the need to generalize the basic geometry of $\Re^2$ and $\Re^3$ to nonvisual higher-dimensional spaces. The usual approach is to coordinatize geometric concepts in $\Re^2$ and $\Re^3$, and then extend statements concerning ordered pairs and triples to ordered n-tuples in $\Re^n$ and $\mathcal{C}^n$.

For example, the length of a vector $\mathbf{u} \in \Re^2$ or $\mathbf{v} \in \Re^3$ is obtained from the Pythagorean theorem by computing the length of the hypotenuse of a right triangle as shown in Figure 5.1.1.

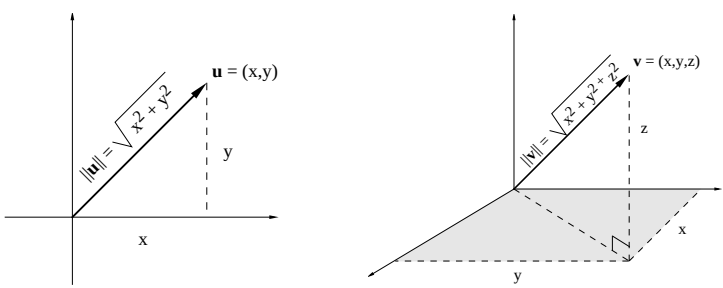


FIGURE 5.1.1

This measure of length,

$$\|\mathbf{u}\| = \sqrt{x^2 + y^2} \quad \text{and} \quad \|\mathbf{v}\| = \sqrt{x^2 + y^2 + z^2},$$

is called the *euclidean norm* in $\Re^2$ and $\Re^3$, and there is an obvious extension to higher dimensions.

Euclidean Vector Norm

For a vector $\mathbf{x}_{n\times 1}$, the ***euclidean norm*** of $\mathbf{x}$ is defined to be

- $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} = \sqrt{\mathbf{x}^T\mathbf{x}}$ whenever $\mathbf{x} \in \Re^{n\times 1}$,
- $\|\mathbf{x}\| = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} = \sqrt{\mathbf{x}^*\mathbf{x}}$ whenever $\mathbf{x} \in \mathcal{C}^{n\times 1}$.

For example, if $\mathbf{u} = \begin{pmatrix} 0 \\ -1 \\ 2 \\ -2 \\ 4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} \mathrm{i} \\ 2 \\ 1-\mathrm{i} \\ 0 \\ 1+\mathrm{i} \end{pmatrix}$, then

$$\|\mathbf{u}\| = \sqrt{\sum u_i^2} = \sqrt{\mathbf{u}^T\mathbf{u}} = \sqrt{0+1+4+4+16} = 5,$$

$$\|\mathbf{v}\| = \sqrt{\sum |v_i|^2} = \sqrt{\mathbf{v}^*\mathbf{v}} = \sqrt{1+4+2+0+2} = 3.$$

There are several points to note.[33]

- The complex version of $\|\mathbf{x}\|$ includes the real version as a special case because $|z|^2 = z^2$ whenever z is a real number. Recall that if $z = a + \mathrm{i}b$, then $\bar{z} = a - \mathrm{i}b$, and the magnitude of z is $|z| = \sqrt{\bar{z}z} = \sqrt{a^2+b^2}$. The fact that $|z|^2 = \bar{z}z = a^2 + b^2$ is a real number insures that $\|\mathbf{x}\|$ is real even if $\mathbf{x}$ has some complex components.
- The definition of euclidean norm guarantees that for all scalars α,

$$\|\mathbf{x}\| \geq 0, \quad \|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \quad \text{and} \quad \|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|. \tag{5.1.1}$$

- Given a vector $\mathbf{x} \neq \mathbf{0}$, it's frequently convenient to have another vector that points in the same direction as $\mathbf{x}$ (i.e., is a positive multiple of $\mathbf{x}$) but has unit length. To construct such a vector, we ***normalize*** $\mathbf{x}$ by setting $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$. From (5.1.1), it's easy to see that

$$\|\mathbf{u}\| = \left\|\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| = \frac{1}{\|\mathbf{x}\|}\|\mathbf{x}\| = 1. \tag{5.1.2}$$

[33] By convention, column vectors are used throughout this chapter. But there is nothing special about columns because, with the appropriate interpretation, all statements concerning columns will also hold for rows.

- The distance between vectors in $\Re^3$ can be visualized with the aid of the parallelogram law as shown in Figure 5.1.2, so for vectors in $\Re^n$ and $\mathcal{C}^n$, the ***distance*** between $\mathbf{u}$ and $\mathbf{v}$ is naturally defined to be $\|\mathbf{u}-\mathbf{v}\|$.

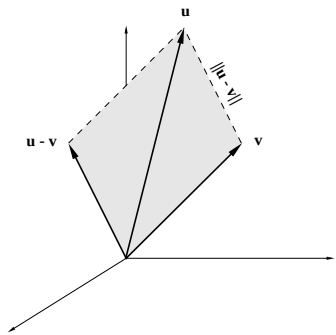


FIGURE 5.1.2

Standard Inner Product

The scalar terms defined by

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i \in \Re \quad \text{and} \quad \mathbf{x}^*\mathbf{y} = \sum_{i=1}^{n} \bar{x}_i y_i \in \mathcal{C}$$

are called the ***standard inner products*** for $\Re^n$ and $\mathcal{C}^n$, respectively.

The Cauchy–Bunyakovskii–Schwarz (CBS) inequality [34] is one of the most important inequalities in mathematics. It relates inner product to norm.

[34] The Cauchy–Bunyakovskii–Schwarz inequality is named in honor of the three men who played a role in its development. The basic inequality for real numbers is attributed to Cauchy in 1821, whereas Schwarz and Bunyakovskii contributed by later formulating useful generalizations of the inequality involving integrals of functions.

Augustin-Louis Cauchy (1789–1857) was a French mathematician who is generally regarded as being the founder of mathematical analysis—including the theory of complex functions. Although deeply embroiled in political turmoil for much of his life (he was a partisan of the Bourbons), Cauchy emerged as one of the most prolific mathematicians of all time. He authored at least 789 mathematical papers, and his collected works fill 27 volumes—this is on a par with Cayley and second only to Euler. It is said that more theorems, concepts, and methods bear Cauchy's name than any other mathematician.

Victor Bunyakovskii (1804–1889) was a Russian professor of mathematics at St. Petersburg, and in 1859 he extended Cauchy's inequality for discrete sums to integrals of continuous functions. His contribution was overlooked by western mathematicians for many years, and his name is often omitted in classical texts that simply refer to the *Cauchy–Schwarz inequality.*

Hermann Amandus Schwarz (1843–1921) was a student and successor of the famous German mathematician Karl Weierstrass at the University of Berlin. Schwarz independently generalized Cauchy's inequality just as Bunyakovskii had done earlier.

Cauchy–Bunyakovskii–Schwarz (CBS) Inequality

$$|\mathbf{x}^*\mathbf{y}| \le \|\mathbf{x}\|\,\|\mathbf{y}\| \quad \text{for all } \mathbf{x},\mathbf{y} \in \mathcal{C}^{n\times 1}. \tag{5.1.3}$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x}$.

Proof. Set $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x} = \mathbf{x}^*\mathbf{y}/\|\mathbf{x}\|^2$ (assume $\mathbf{x} \neq \mathbf{0}$ because there is nothing to prove if $\mathbf{x} = \mathbf{0}$) and observe that $\mathbf{x}^*(\alpha\mathbf{x} - \mathbf{y}) = 0$, so

$$\begin{aligned} 0 &\le \|\alpha\mathbf{x} - \mathbf{y}\|^2 = (\alpha\mathbf{x} - \mathbf{y})^*(\alpha\mathbf{x} - \mathbf{y}) = \bar{\alpha}\mathbf{x}^*(\alpha\mathbf{x} - \mathbf{y}) - \mathbf{y}^*(\alpha\mathbf{x} - \mathbf{y}) \\ &= -\mathbf{y}^*(\alpha\mathbf{x} - \mathbf{y}) = \mathbf{y}^*\mathbf{y} - \alpha\mathbf{y}^*\mathbf{x} = \frac{\|\mathbf{y}\|^2\,\|\mathbf{x}\|^2 - (\mathbf{x}^*\mathbf{y})\,(\mathbf{y}^*\mathbf{x})}{\|\mathbf{x}\|^2}. \end{aligned} \tag{5.1.4}$$

Since $\mathbf{y}^*\mathbf{x} = \overline{\mathbf{x}^*\mathbf{y}}$, it follows that $(\mathbf{x}^*\mathbf{y})\,(\mathbf{y}^*\mathbf{x}) = |\mathbf{x}^*\mathbf{y}|^2$, so

$$0 \le \frac{\|\mathbf{y}\|^2\,\|\mathbf{x}\|^2 - |\mathbf{x}^*\mathbf{y}|^2}{\|\mathbf{x}\|^2}.$$

Now, $0 < \|\mathbf{x}\|^2$ implies $0 \le \|\mathbf{y}\|^2\,\|\mathbf{x}\|^2 - |\mathbf{x}^*\mathbf{y}|^2$, and thus the CBS inequality is obtained. Establishing the conditions for equality is Exercise 5.1.9. ∎

One reason that the CBS inequality is important is because it helps to establish that the geometry in higher-dimensional spaces is consistent with the geometry in the visual spaces $\Re^2$ and $\Re^3$. In particular, consider the situation depicted in Figure 5.1.3.

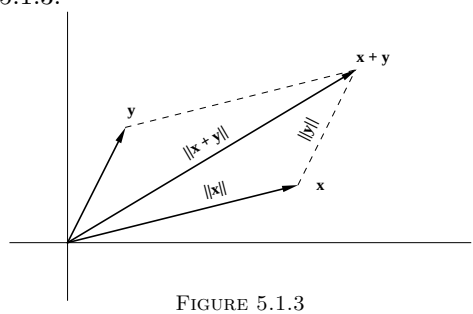


Figure 5.1.3

Imagine traveling from the origin to the point $\mathbf{x}$ and then moving from $\mathbf{x}$ to the point $\mathbf{x}+\mathbf{y}$. Clearly, you have traveled a distance that is at least as great as the direct distance from the origin to $\mathbf{x}+\mathbf{y}$ along the diagonal of the parallelogram. In other words, it's visually evident that $\|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$. This observation

is known as the ***triangle inequality.*** In higher-dimensional spaces we do not have the luxury of visualizing the geometry with our eyes, and the question of whether or not the triangle inequality remains valid has no obvious answer. The CBS inequality is precisely what is required to prove that, in this respect, the geometry of higher dimensions is no different than that of the visual spaces.

Triangle Inequality

$$\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad \text{for every } \mathbf{x}, \mathbf{y} \in \mathcal{C}^n.$$

Proof. Consider $\mathbf{x}$ and $\mathbf{y}$ to be column vectors, and write

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|^2 &= (\mathbf{x}+\mathbf{y})^*(\mathbf{x}+\mathbf{y}) = \mathbf{x}^*\mathbf{x} + \mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} + \mathbf{y}^*\mathbf{y} \\ &= \|\mathbf{x}\|^2 + \mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} + \|\mathbf{y}\|^2 . \end{aligned} \tag{5.1.5}$$

Recall that if $z = a + \mathrm{i}b$, then $z + \bar{z} = 2a = 2\,\mathrm{Re}\,(z)$ and $|z|^2 = a^2 + b^2 \geq a^2$, so that $|z| \geq \mathrm{Re}\,(z)$. Using the fact that $\mathbf{y}^*\mathbf{x} = \overline{\mathbf{x}^*\mathbf{y}}$ together with the CBS inequality yields

$$\mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} = 2\,\mathrm{Re}\,(\mathbf{x}^*\mathbf{y}) \leq 2\,|\mathbf{x}^*\mathbf{y}| \leq 2\,\|\mathbf{x}\|\,\|\mathbf{y}\| .$$

Consequently, we may infer from (5.1.5) that

$$\|\mathbf{x}+\mathbf{y}\|^2 \leq \|\mathbf{x}\|^2 + 2\,\|\mathbf{x}\|\,\|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 . \quad \blacksquare$$

It's not difficult to see that the triangle inequality can be extended to any number of vectors in the sense that $\left\|\sum_i \mathbf{x}_i\right\| \leq \sum_i \|\mathbf{x}_i\|$. Furthermore, it follows as a corollary that for real or complex numbers, $\left|\sum_i \alpha_i\right| \leq \sum_i |\alpha_i|$ (the triangle inequality for scalars).

Example 5.1.1

Backward Triangle Inequality. The triangle inequality produces an upper bound for a sum, but it also yields the following lower bound for a difference:

$$\big|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,\big| \leq \|\mathbf{x}-\mathbf{y}\| . \tag{5.1.6}$$

This is a consequence of the triangle inequality because

$$\|\mathbf{x}\| = \|\mathbf{x}-\mathbf{y}+\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{y}\| \implies \|\mathbf{x}\| - \|\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\|$$

and

$$\|\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}-\mathbf{x}\| \leq \|\mathbf{x}-\mathbf{y}\| + \|\mathbf{x}\| \implies -(\|\mathbf{x}\| - \|\mathbf{y}\|) \leq \|\mathbf{x}-\mathbf{y}\| .$$

There are notions of length other than the euclidean measure. For example, urban dwellers navigate on a grid of city blocks with one-way streets, so they are prone to measure distances in the city not as the crow flies but rather in terms of lengths on a directed grid. For example, instead of than saying that "it's a one-half mile straight-line (euclidean) trip from here to there," they are more apt to describe the length of the trip by saying, "it's two blocks north on Dan Allen Drive, four blocks west on Hillsborough Street, and five blocks south on Gorman Street." In other words, the length of the trip is $2 + |-4| + |-5| = 11$ blocks—absolute value is used to insure that southerly and westerly movement does not cancel the effect of northerly and easterly movement, respectively. This "grid norm" is better known as the 1-norm because it is a special case of a more general class of norms defined below.

p-Norms

For $p \geq 1$, the ***p-norm*** of $\mathbf{x} \in \mathcal{C}^n$ is defined as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$.

It can be proven that the following properties of the euclidean norm are in fact valid for all p-norms:

$$\begin{aligned} &\|\mathbf{x}\|_p \geq 0 \quad \text{and} \quad \|\mathbf{x}\|_p = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \\ &\|\alpha\mathbf{x}\|_p = |\alpha|\ \|\mathbf{x}\|_p \quad \text{for all scalars } \alpha, \\ &\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p \qquad \text{(see Exercise 5.1.13).} \end{aligned} \tag{5.1.7}$$

The generalized version of the CBS inequality (5.1.3) for p-norms is ***Hölder's inequality*** (developed in Exercise 5.1.12), which states that if $p > 1$ and $q > 1$ are real numbers such that $1/p + 1/q = 1$, then

$$|\mathbf{x}^*\mathbf{y}| \leq \|\mathbf{x}\|_p\ \|\mathbf{y}\|_q\,. \tag{5.1.8}$$

In practice, only three of the p-norms are used, and they are

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| \quad \text{(the grid norm)}, \quad \|\mathbf{x}\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2} \text{(the euclidean norm)},$$

and

$$\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p = \lim_{p\to\infty} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} = \max_i |x_i| \qquad \text{(the max norm)}.$$

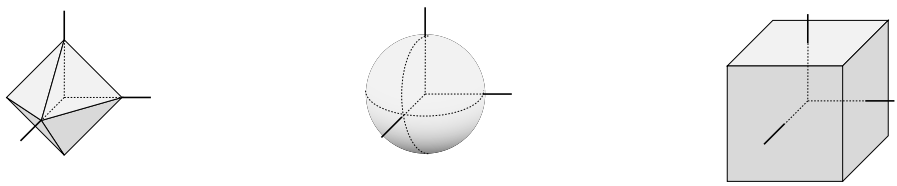
For example, if $\mathbf{x} = (3,\ 4-3\mathrm{i},\ 1)$, then $\|\mathbf{x}\|_1 = 9$, $\|\mathbf{x}\|_2 = \sqrt{35}$, and $\|\mathbf{x}\|_\infty = 5$.

To see that $\lim_{p\to\infty} \|\mathbf{x}\|_p = \max_i |x_i|$, proceed as follows. Relabel the entries of $\mathbf{x}$ by setting $\tilde{x}_1 = \max_i |x_i|$, and if there are other entries with this same maximal magnitude, label them $\tilde{x}_2, \ldots, \tilde{x}_k$. Label any remaining coordinates as $\tilde{x}_{k+1} \cdots \tilde{x}_n$. Consequently, $|\tilde{x}_i/\tilde{x}_1| < 1$ for $i = k+1, \ldots, n$, so, as $p \to \infty$,

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\tilde{x}_i|^p\right)^{1/p} = |\tilde{x}_1| \left(k + \left|\frac{\tilde{x}_{k+1}}{\tilde{x}_1}\right|^p + \cdots + \left|\frac{\tilde{x}_n}{\tilde{x}_1}\right|^p\right)^{1/p} \to |\tilde{x}_1|\,.$$

Example 5.1.2

To get a feel for the 1-, 2-, and ∞-norms, it helps to know the shapes and relative sizes of the ***unit p-spheres*** $\mathcal{S}_p = \{\mathbf{x} \mid \|\mathbf{x}\|_p = 1\}$ for $p = 1, 2, \infty$. As illustrated in Figure 5.1.4, the unit 1-, 2-, and ∞-spheres in $\Re^3$ are an octahedron, a ball, and a cube, respectively, and it's visually evident that $\mathcal{S}_1$ fits inside $\mathcal{S}_2$, which in turn fits inside $\mathcal{S}_\infty$. This means that $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \Re^3$. In general, this is true in $\Re^n$ (Exercise 5.1.8).



$\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_\infty$

FIGURE 5.1.4

Because the p-norms are defined in terms of coordinates, their use is limited to coordinate spaces. But it's desirable to have a general notion of norm that works for *all* vector spaces. In other words, we need a coordinate-free definition of norm that includes the standard p-norms as a special case. Since all of the p-norms satisfy the properties (5.1.7), it's natural to use these properties to extend the concept of norm to general vector spaces.

General Vector Norms

A ***norm*** for a real or complex vector space $\mathcal{V}$ is a function $\|\star\|$ mapping $\mathcal{V}$ into $\Re$ that satisfies the following conditions.

$$\begin{aligned} &\|\mathbf{x}\| \geq 0 \quad \text{and} \quad \|\mathbf{x}\| = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}, \\ &\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\| \quad \text{for all scalars } \alpha, \\ &\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|\,. \end{aligned} \tag{5.1.9}$$

Example 5.1.3

Equivalent Norms. Vector norms are basic tools for defining and analyzing limiting behavior in vector spaces $\mathcal{V}$. A sequence $\{\mathbf{x}_k\} \subset \mathcal{V}$ is said to converge to $\mathbf{x}$ (write $\mathbf{x}_k \to \mathbf{x}$) if $\|\mathbf{x}_k - \mathbf{x}\| \to 0$. This depends on the choice of the norm, so, ostensibly, we might have $\mathbf{x}_k \to \mathbf{x}$ with one norm but not with another. Fortunately, this is impossible in finite-dimensional spaces because all norms are *equivalent* in the following sense.

Problem: For each pair of norms, $\|\star\|_a$, $\|\star\|_b$, on an n-dimensional space $\mathcal{V}$, exhibit positive constants α and β (depending only on the norms) such that

$$\alpha \le \frac{\|\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \le \beta \quad \text{for all nonzero vectors in } \mathcal{V}. \tag{5.1.10}$$

Solution: For $\mathcal{S}_b = \{\mathbf{y} \mid \|\mathbf{y}\|_b = 1\}$, let $\mu = \min_{\mathbf{y}\in\mathcal{S}_b} \|\mathbf{y}\|_a > 0$, [35] and write

$$\frac{\mathbf{x}}{\|\mathbf{x}\|_b} \in \mathcal{S}_b \quad \Longrightarrow \quad \|\mathbf{x}\|_a = \|\mathbf{x}\|_b \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_b} \right\|_a \ge \|\mathbf{x}\|_b \min_{\mathbf{y}\in\mathcal{S}_b} \|\mathbf{y}\|_a = \|\mathbf{x}\|_b\, \mu.$$

The same argument shows there is a $\nu > 0$ such that $\|\mathbf{x}\|_b \ge \nu \|\mathbf{x}\|_a$, so (5.1.10) is produced with $\alpha = \mu$ and $\beta = 1/\nu$. Note that (5.1.10) insures that $\|\mathbf{x}_k - \mathbf{x}\|_a \to 0$ if and only if $\|\mathbf{x}_k - \mathbf{x}\|_b \to 0$. Specific values for α and β are given in Exercises 5.1.8 and 5.12.3.

Exercises for section 5.1

5.1.1. Find the 1-, 2-, and ∞-norms of $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \\ -4 \\ -2 \end{pmatrix}$ and $\mathbf{x} = \begin{pmatrix} 1+\mathrm{i} \\ 1-\mathrm{i} \\ 1 \\ 4\mathrm{i} \end{pmatrix}$.

5.1.2. Consider the euclidean norm with $\mathbf{u} = \begin{pmatrix} 2 \\ 1 \\ -4 \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$.

(a) Determine the distance between $\mathbf{u}$ and $\mathbf{v}$.
(b) Verify that the triangle inequality holds for $\mathbf{u}$ and $\mathbf{v}$.
(c) Verify that the CBS inequality holds for $\mathbf{u}$ and $\mathbf{v}$.

5.1.3. Show that $(\alpha_1 + \alpha_2 + \cdots + \alpha_n)^2 \le n\left(\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2\right)$ for $\alpha_i \in \Re$.

[35] An important theorem from analysis states that a continuous function mapping a closed and bounded subset $\mathcal{K} \subset \mathcal{V}$ into $\Re$ attains a minimum and maximum value at points in $\mathcal{K}$. Unit spheres in finite-dimensional spaces are closed and bounded, and every norm on $\mathcal{V}$ is continuous (Exercise 5.1.7), so this minimum is guaranteed to exist.

5.1.4. (a) Using the euclidean norm, describe the solid ball in $\Re^n$ centered at the origin with unit radius. (b) Describe a solid ball centered at the point $\mathbf{c} = (\,\xi_1 \quad \xi_2 \quad \cdots \quad \xi_n\,)$ with radius ρ.

5.1.5. If $\mathbf{x}, \mathbf{y} \in \Re^n$ such that $\|\mathbf{x} - \mathbf{y}\|_2 = \|\mathbf{x} + \mathbf{y}\|_2$, what is $\mathbf{x}^T\mathbf{y}$?

5.1.6. Explain why $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\|$ is true for all norms.

5.1.7. For every vector norm on $\mathcal{C}^n$, prove that $\|\mathbf{v}\|$ depends continuously on the components of $\mathbf{v}$ in the sense that for each $\epsilon > 0$, there corresponds a $\delta > 0$ such that $\big|\,\|\mathbf{x}\| - \|\mathbf{y}\|\,\big| < \epsilon$ whenever $|x_i - y_i| < \delta$ for each i.

5.1.8. (a) For $\mathbf{x} \in \mathcal{C}^{n\times 1}$, explain why $\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_\infty$.

(b) For $\mathbf{x} \in \mathcal{C}^{n\times 1}$, show that $\|\mathbf{x}\|_i \leq \alpha\,\|\mathbf{x}\|_j$, where α is the (i,j)-entry in the following matrix. (See Exercise 5.12.3 for a similar statement regarding matrix norms.)

$$\begin{array}{c} \\ 1 \\ 2 \\ \infty \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & \infty \end{array} \\ \begin{pmatrix} * & \sqrt{n} & n \\ 1 & * & \sqrt{n} \\ 1 & 1 & * \end{pmatrix} \end{array}.$$

5.1.9. For $\mathbf{x}, \mathbf{y} \in \mathcal{C}^n$, $\mathbf{x} \neq \mathbf{0}$, explain why equality holds in the CBS inequality if and only if $\mathbf{y} = \alpha\mathbf{x}$, where $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x}$. **Hint:** Use (5.1.4).

5.1.10. For nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathcal{C}^n$ with the euclidean norm, prove that equality holds in the triangle inequality if and only if $\mathbf{y} = \alpha\mathbf{x}$, where α is real and positive. **Hint:** Make use of Exercise 5.1.9.

5.1.11. Use Hölder's inequality (5.1.8) to prove that if the components of $\mathbf{x} \in \Re^{n\times 1}$ sum to zero (i.e., $\mathbf{x}^T\mathbf{e} = 0$ for $\mathbf{e}^T = (1, 1, \ldots, 1)$), then

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_1 \left(\frac{y_{\max} - y_{\min}}{2}\right) \quad \text{for all } \mathbf{y} \in \Re^{n\times 1}.$$

Note: For "zero sum" vectors $\mathbf{x}$, this is at least as sharp and usually it's sharper than (5.1.8) because $(y_{\max} - y_{\min})/2 \leq \max_i |y_i| = \|\mathbf{y}\|_\infty$.

5.1.12. The classical form of ***Hölder's inequality***[36] states that if $p > 1$ and $q > 1$ are real numbers such that $1/p + 1/q = 1$, then

$$\sum_{i=1}^{n} |x_i y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q \right)^{1/q}.$$

Derive this inequality by executing the following steps:

(a) By considering the function $f(t) = (1-\lambda) + \lambda t - t^\lambda$ for $0 < \lambda < 1$, establish the inequality

$$\alpha^\lambda \beta^{1-\lambda} \leq \lambda\alpha + (1-\lambda)\beta$$

for nonnegative real numbers α and β.

(b) Let $\hat{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_p$ and $\hat{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_q$, and apply the inequality of part (a) to obtain

$$\sum_{i=1}^{n} |\hat{x}_i \hat{y}_i| \leq \frac{1}{p} \sum_{i=1}^{n} |\hat{x}_i|^p + \frac{1}{q} \sum_{i=1}^{n} |\hat{y}_i|^q = 1.$$

(c) Deduce the classical form of Hölder's inequality, and then explain why this means that

$$|\mathbf{x}^* \mathbf{y}| \leq \|\mathbf{x}\|_p \, \|\mathbf{y}\|_q .$$

5.1.13. The triangle inequality $\|\mathbf{x}+\mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$ for a general p-norm is really the classical ***Minkowski inequality,***[37] which states that for $p \geq 1$,

$$\left(\sum_{i=1}^{n} |x_i + y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p \right)^{1/p}.$$

Derive Minkowski's inequality. **Hint:** For $p > 1$, let q be the number such that $1/q = 1 - 1/p$. Verify that for scalars α and β,

$$|\alpha+\beta|^p = |\alpha+\beta|\,|\alpha+\beta|^{p/q} \leq |\alpha|\,|\alpha+\beta|^{p/q} + |\beta|\,|\alpha+\beta|^{p/q},$$

and make use of Hölder's inequality in Exercise 5.1.12.

---

[36] Ludwig Otto Hölder (1859–1937) was a German mathematician who studied at Göttingen and lived in Leipzig. Although he made several contributions to analysis as well as algebra, he is primarily known for the development of the inequality that now bears his name.

[37] Hermann Minkowski (1864–1909) was born in Russia, but spent most of his life in Germany as a mathematician and professor at Königsberg and Göttingen. In addition to the inequality that now bears his name, he is known for providing a mathematical basis for the special theory of relativity. He died suddenly from a ruptured appendix at the age of 44.

5.2 MATRIX NORMS

Because $\mathcal{C}^{m\times n}$ is a vector space of dimension mn, magnitudes of matrices $\mathbf{A} \in \mathcal{C}^{m\times n}$ can be "measured" by employing any vector norm on $\mathcal{C}^{mn}$. For example, by stringing out the entries of $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -4 & -2 \end{pmatrix}$ into a four-component vector, the euclidean norm on $\Re^4$ can be applied to write

$$\|\mathbf{A}\| = \left[2^2 + (-1)^2 + (-4)^2 + (-2)^2\right]^{1/2} = 5.$$

This is one of the simplest notions of a matrix norm, and it is called the *Frobenius* (p. 662) *norm* (older texts refer to it as the *Hilbert–Schmidt norm* or the *Schur norm*). There are several useful ways to describe the Frobenius matrix norm.

Frobenius Matrix Norm

The ***Frobenius norm*** of $\mathbf{A} \in \mathcal{C}^{m\times n}$ is defined by the equations

$$\|\mathbf{A}\|_F^2 = \sum_{i,j} |a_{ij}|^2 = \sum_i \|\mathbf{A}_{i*}\|_2^2 = \sum_j \|\mathbf{A}_{*j}\|_2^2 = trace\,(\mathbf{A}^*\mathbf{A}). \tag{5.2.1}$$

The Frobenius matrix norm is fine for some problems, but it is not well suited for all applications. So, similar to the situation for vector norms, alternatives need to be explored. But before trying to develop different recipes for matrix norms, it makes sense to first formulate a general definition of a matrix norm. The goal is to start with the defining properties for a vector norm given in (5.1.9) on p. 275 and ask what, if anything, needs to be added to that list.

Matrix multiplication distinguishes matrix spaces from more general vector spaces, but the three vector-norm properties (5.1.9) say nothing about products. So, an extra property that relates $\|\mathbf{AB}\|$ to $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ is needed. The Frobenius norm suggests the nature of this extra property. The CBS inequality insures that $\|\mathbf{Ax}\|_2^2 = \sum_i |\mathbf{A}_{i*}\mathbf{x}|^2 \le \sum_i \|\mathbf{A}_{i*}\|_2^2 \|\mathbf{x}\|_2^2 = \|\mathbf{A}\|_F^2 \|\mathbf{x}\|_2^2$. That is,

$$\|\mathbf{Ax}\|_2 \le \|\mathbf{A}\|_F \|\mathbf{x}\|_2, \tag{5.2.2}$$

and we express this by saying that the Frobenius matrix norm $\|\star\|_F$ and the euclidean vector norm $\|\star\|_2$ are ***compatible***. The compatibility condition (5.2.2) implies that for all conformable matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\begin{aligned}\|\mathbf{AB}\|_F^2 &= \sum_j \|[\mathbf{AB}]_{*j}\|_2^2 = \sum_j \|\mathbf{AB}_{*j}\|_2^2 \le \sum_j \|\mathbf{A}\|_F^2 \|\mathbf{B}_{*j}\|_2^2 \\ &= \|\mathbf{A}\|_F^2 \sum_j \|\mathbf{B}_{*j}\|_2^2 = \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \implies \|\mathbf{AB}\|_F \le \|\mathbf{A}\|_F \|\mathbf{B}\|_F.\end{aligned}$$

This suggests that the submultiplicative property $\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|$ should be added to (5.1.9) to define a general matrix norm.

General Matrix Norms

A ***matrix norm*** is a function $\|\star\|$ from the set of all complex matrices (of all finite orders) into $\Re$ that satisfies the following properties.

$$\begin{aligned}
&\|\mathbf{A}\| \geq 0 \quad \text{and} \quad \|\mathbf{A}\| = 0 \Longleftrightarrow \mathbf{A} = \mathbf{0}.\\
&\|\alpha\mathbf{A}\| = |\alpha|\,\|\mathbf{A}\| \quad \text{for all scalars } \alpha.\\
&\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \quad \text{for matrices of the same size.}\\
&\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\| \quad \text{for all conformable matrices.}
\end{aligned} \tag{5.2.3}$$

The Frobenius norm satisfies the above definition (it was built that way), but where do other useful matrix norms come from? In fact, every legitimate vector norm generates (or induces) a matrix norm as described below.

Induced Matrix Norms

A vector norm that is defined on $\mathcal{C}^p$ for $p = m, n$ ***induces*** a matrix norm on $\mathcal{C}^{m\times n}$ by setting

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\| \quad \text{for } \mathbf{A} \in \mathcal{C}^{m\times n},\ \ \mathbf{x} \in \mathcal{C}^{n\times 1}. \tag{5.2.4}$$

The footnote on p. 276 explains why this maximum value must exist.

- It's apparent that an induced matrix norm is compatible with its underlying vector norm in the sense that

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|\,. \tag{5.2.5}$$

- When $\mathbf{A}$ is nonsingular, $\displaystyle\min_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\| = \frac{1}{\|\mathbf{A}^{-1}\|}$. (5.2.6)

Proof. Verifying that $\max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$ satisfies the first three conditions in (5.2.3) is straightforward, and (5.2.5) implies $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$ (see Exercise 5.2.5). Property (5.2.6) is developed in Exercise 5.2.7. ∎

In words, an induced norm $\|\mathbf{A}\|$ represents the maximum extent to which a vector on the unit sphere can be stretched by $\mathbf{A}$, and $1/\left\|\mathbf{A}^{-1}\right\|$ measures the extent to which a nonsingular matrix $\mathbf{A}$ can shrink vectors on the unit sphere. Figure 5.2.1 depicts this in $\Re^3$ for the induced matrix 2-norm.

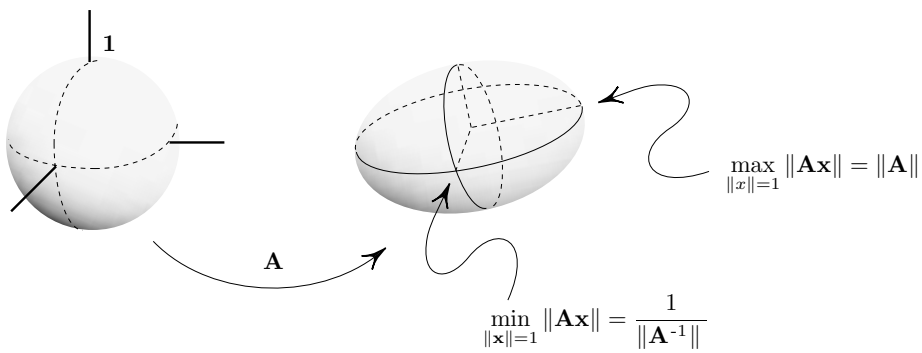


FIGURE 5.2.1. THE INDUCED MATRIX 2-NORM IN $\Re^3$.

Intuition might suggest that the euclidean vector norm should induce the Frobenius matrix norm (5.2.1), but something surprising happens instead.

Matrix 2-Norm

- The matrix norm induced by the euclidean vector norm is

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \sqrt{\lambda_{\max}}, \tag{5.2.7}$$

  where $\lambda_{\max}$ is the largest number λ such that $\mathbf{A}^*\mathbf{A} - \lambda\mathbf{I}$ is singular.

- When $\mathbf{A}$ is nonsingular,

$$\left\|\mathbf{A}^{-1}\right\|_2 = \frac{1}{\min\limits_{\|x\|_2=1} \|\mathbf{Ax}\|_2} = \frac{1}{\sqrt{\lambda_{\min}}}, \tag{5.2.8}$$

  where $\lambda_{\min}$ is the smallest number λ such that $\mathbf{A}^*\mathbf{A} - \lambda\mathbf{I}$ is singular.

Note: If you are already familiar with eigenvalues, these say that $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\mathbf{A}^*\mathbf{A}$ (Example 7.5.1, p. 549), while $(\lambda_{\max})^{1/2} = \sigma_1$ and $(\lambda_{\min})^{1/2} = \sigma_n$ are the largest and smallest singular values of $\mathbf{A}$ (p. 414).

Proof. To prove (5.2.7), assume that $\mathbf{A}_{m\times n}$ is real (a proof for complex matrices is given in Example 7.5.1 on p. 549). The strategy is to evaluate $\|\mathbf{A}\|_2^2$ by solving the problem

$$\text{maximize } f(\mathbf{x}) = \|\mathbf{Ax}\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} \quad \text{subject to } g(\mathbf{x}) = \mathbf{x}^T\mathbf{x} = 1$$

using the method of Lagrange multipliers. Introduce a new variable λ (the Lagrange multiplier), and consider the function $h(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$. The points at which f is maximized are contained in the set of solutions to the equations $\partial h/\partial x_i = 0$ $(i = 1, 2, \ldots, n)$ along with $g(\mathbf{x}) = 1$. Differentiating h with respect to the x_i's is essentially the same as described on p. 227, and the system generated by $\partial h/\partial x_i = 0$ $(i = 1, 2, \ldots, n)$ is $(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$. In other words, f is maximized at a vector $\mathbf{x}$ for which $(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ and $\|\mathbf{x}\|_2 = 1$. Consequently, λ must be a number such that $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular (because $\mathbf{x} \neq \mathbf{0}$). Since

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda,$$

it follows that

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \max_{\|\mathbf{x}\|^2=1} \|\mathbf{A}\mathbf{x}\| = \left(\max_{\mathbf{x}^T\mathbf{x}=1} \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}\right)^{1/2} = \sqrt{\lambda_{\max}},$$

where $\lambda_{\max}$ is the largest number λ for which $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular. A similar argument applied to (5.2.6) proves (5.2.8). Also, an independent development of (5.2.7) and (5.2.8) is contained in the discussion of singular values on p. 412. ∎

Example 5.2.1

Problem: Determine the induced norm $\|\mathbf{A}\|_2$ as well as $\|\mathbf{A}^{-1}\|_2$ for the nonsingular matrix

$$\mathbf{A} = \frac{1}{\sqrt{3}}\begin{pmatrix} 3 & -1 \\ 0 & \sqrt{8} \end{pmatrix}.$$

Solution: Find the values of λ that make $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ singular by applying Gaussian elimination to produce

$$\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{pmatrix} \longrightarrow \begin{pmatrix} -1 & 3-\lambda \\ 3-\lambda & -1 \end{pmatrix} \longrightarrow \begin{pmatrix} -1 & 3-\lambda \\ 0 & -1+(3-\lambda)^2 \end{pmatrix}.$$

This shows that $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular when $-1 + (3-\lambda)^2 = 0$ or, equivalently, when $\lambda = 2$ or $\lambda = 4$, so $\lambda_{\min} = 2$ and $\lambda_{\max} = 4$. Consequently, (5.2.7) and (5.2.8) say that

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}} = 2 \quad \text{and} \quad \|\mathbf{A}^{-1}\|_2 = \frac{1}{\sqrt{\lambda_{\min}}} = \frac{1}{\sqrt{2}}.$$

Note: As mentioned earlier, the values of λ that make $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ singular are called the *eigenvalues* of $\mathbf{A}^T\mathbf{A}$, and they are the focus of Chapter 7 where their determination is discussed in more detail. Using Gaussian elimination to determine the eigenvalues is not practical for larger matrices.

Some useful properties of the matrix 2-norm are stated below.

Properties of the 2-Norm

In addition to the properties shared by all induced norms, the 2-norm enjoys the following special properties.

- $$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \max_{\|\mathbf{y}\|_2=1} |\mathbf{y}^*\mathbf{A}\mathbf{x}|. \tag{5.2.9}$$
- $$\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2. \tag{5.2.10}$$
- $$\|\mathbf{A}^*\mathbf{A}\|_2 = \|\mathbf{A}\|_2^2. \tag{5.2.11}$$
- $$\left\| \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix} \right\|_2 = \max\{\, \|\mathbf{A}\|_2 \,,\, \|\mathbf{B}\|_2 \,\}. \tag{5.2.12}$$
- $$\|\mathbf{U}^*\mathbf{A}\mathbf{V}\|_2 = \|\mathbf{A}\|_2 \text{ when } \mathbf{U}\mathbf{U}^* = \mathbf{I} \text{ and } \mathbf{V}^*\mathbf{V} = \mathbf{I}. \tag{5.2.13}$$

You are asked to verify the validity of these properties in Exercise 5.2.6 on p. 285. Furthermore, some additional properties of the matrix 2-norm are developed in Exercise 5.6.9 and on pp. 414 and 417.

Now that we understand how the euclidean vector norm induces the matrix 2-norm, let's investigate the nature of the matrix norms that are induced by the vector 1-norm and the vector ∞-norm.

Matrix 1-Norm and Matrix ∞-Norm

The matrix norms induced by the vector 1-norm and ∞-norm are as follows.

- $$\|\mathbf{A}\|_1 = \max_{\|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1 = \max_j \sum_i |a_{ij}| \tag{5.2.14}$$
  $$= \text{the largest absolute column sum.}$$
- $$\|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_i \sum_j |a_{ij}| \tag{5.2.15}$$
  $$= \text{the largest absolute row sum.}$$

Proof of (5.2.14). For all $\mathbf{x}$ with $\|\mathbf{x}\|_1 = 1$, the scalar triangle inequality yields

$$\|\mathbf{A}\mathbf{x}\|_1 = \sum_i |\mathbf{A}_{i*}\mathbf{x}| = \sum_i \Big| \sum_j a_{ij}x_j \Big| \le \sum_i \sum_j |a_{ij}|\,|x_j| = \sum_j \Big(|x_j| \sum_i |a_{ij}| \Big)$$
$$\le \Big(\sum_j |x_j| \Big) \Big(\max_j \sum_i |a_{ij}| \Big) = \max_j \sum_i |a_{ij}|\,.$$

Equality can be attained because if $\mathbf{A}_{*k}$ is the column with largest absolute sum, set $\mathbf{x} = \mathbf{e}_k$, and note that $\|\mathbf{e}_k\|_1 = 1$ and $\|\mathbf{A}\mathbf{e}_k\|_1 = \|\mathbf{A}_{*k}\|_1 = \max_j \sum_i |a_{ij}|$.

Proof of (5.2.15). For all $\mathbf{x}$ with $\|\mathbf{x}\|_\infty = 1$,

$$\|\mathbf{A}\mathbf{x}\|_\infty = \max_i \left| \sum_j a_{ij} x_j \right| \le \max_i \sum_j |a_{ij}|\ |x_j| \le \max_i \sum_j |a_{ij}|\,.$$

Equality can be attained because if $\mathbf{A}_{k*}$ is the row with largest absolute sum, and if $\mathbf{x}$ is the vector such that

$$x_j = \begin{cases} 1 & \text{if } a_{kj} \ge 0, \\ -1 & \text{if } a_{kj} < 0, \end{cases} \quad \text{then} \quad \begin{cases} |\mathbf{A}_{i*}\mathbf{x}| = |\sum_j a_{ij} x_j| \le \sum_j |a_{ij}| \text{ for all } i, \\ |\mathbf{A}_{k*}\mathbf{x}| = \sum_j |a_{kj}| = \max_i \sum_j |a_{ij}|\,, \end{cases}$$

so $\|\mathbf{x}\|_\infty = 1$, and $\|\mathbf{A}\mathbf{x}\|_\infty = \max_i |\mathbf{A}_{i*}\mathbf{x}| = \max_i \sum_j |a_{ij}|$. ∎

Example 5.2.2

Problem: Determine the induced matrix norms $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$ for

$$\mathbf{A} = \frac{1}{\sqrt{3}} \begin{pmatrix} 3 & -1 \\ 0 & \sqrt{8} \end{pmatrix},$$

and compare the results with $\|\mathbf{A}\|_2$ (from Example 5.2.1) and $\|\mathbf{A}\|_F$.

Solution: Equation (5.2.14) says that $\|\mathbf{A}\|_1$ is the largest absolute column sum in $\mathbf{A}$, and (5.2.15) says that $\|\mathbf{A}\|_\infty$ is the largest absolute row sum, so

$$\|\mathbf{A}\|_1 = 1/\sqrt{3} + \sqrt{8}/\sqrt{3} \approx 2.21 \quad \text{and} \quad \|\mathbf{A}\|_\infty = 4/\sqrt{3} \approx 2.31.$$

Since $\|\mathbf{A}\|_2 = 2$ (Example 5.2.1) and $\|\mathbf{A}\|_F = \sqrt{trace\,(\mathbf{A}^T\mathbf{A})} = \sqrt{6} \approx 2.45$, we see that while $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|_F$ are not equal, they are all in the same ballpark. This is true for all $n \times n$ matrices because it can be shown that $\|\mathbf{A}\|_i \le \alpha \|\mathbf{A}\|_j$, where α is the (i,j)-entry in the following matrix

$$\begin{array}{c} \\ 1 \\ 2 \\ \infty \\ F \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \infty & F \end{array} \\ \begin{pmatrix} * & \sqrt{n} & n & \sqrt{n} \\ \sqrt{n} & * & \sqrt{n} & 1 \\ n & \sqrt{n} & * & \sqrt{n} \\ \sqrt{n} & \sqrt{n} & \sqrt{n} & * \end{pmatrix} \end{array}$$

(see Exercise 5.1.8 and Exercise 5.12.3 on p. 425). Since it's often the case that only the order of magnitude of $\|\mathbf{A}\|$ is needed and not the exact value (e.g., recall the rule of thumb in Example 3.8.2 on p. 129), and since $\|\mathbf{A}\|_2$ is difficult to compute in comparison with $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, and $\|\mathbf{A}\|_F$, you can see why any of these three might be preferred over $\|\mathbf{A}\|_2$ in spite of the fact that $\|\mathbf{A}\|_2$ is more "natural" by virtue of being induced by the euclidean vector norm.

Exercises for section 5.2

5.2.1. Evaluate the Frobenius matrix norm for each matrix below.

$$\mathbf{A} = \begin{pmatrix} 1 & -2 \\ -1 & 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 4 & -2 & 4 \\ -2 & 1 & -2 \\ 4 & -2 & 4 \end{pmatrix}.$$

5.2.2. Evaluate the induced 1-, 2-, and ∞-matrix norm for each of the three matrices given in Exercise 5.2.1.

5.2.3. (a) Explain why $\|\mathbf{I}\| = 1$ for every induced matrix norm (5.2.4).
(b) What is $\|\mathbf{I}_{n\times n}\|_F$?

5.2.4. Explain why $\|\mathbf{A}\|_F = \|\mathbf{A}^*\|_F$ for Frobenius matrix norm (5.2.1).

5.2.5. For matrices $\mathbf{A}$ and $\mathbf{B}$ and for vectors $\mathbf{x}$, establish the following compatibility properties between a vector norm defined on every $\mathcal{C}^p$ and the associated induced matrix norm.
(a) Show that $\|\mathbf{Ax}\| \le \|\mathbf{A}\|\,\|\mathbf{x}\|$.
(b) Show that $\|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|$.
(c) Explain why $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|\le 1} \|\mathbf{Ax}\|$.

5.2.6. Establish the following properties of the matrix 2-norm.
(a) $\|\mathbf{A}\|_2 = \max\limits_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{y}^*\mathbf{Ax}|$,
(b) $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2$,
(c) $\|\mathbf{A}^*\mathbf{A}\|_2 = \|\mathbf{A}\|_2^2$,
(d) $\left\|\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}\right\|_2 = \max\{\,\|\mathbf{A}\|_2\,,\ \|\mathbf{B}\|_2\,\}$ (take $\mathbf{A}$, $\mathbf{B}$ to be real),
(e) $\|\mathbf{U}^*\mathbf{AV}\|_2 = \|\mathbf{A}\|_2$ when $\mathbf{UU}^* = \mathbf{I}$ and $\mathbf{V}^*\mathbf{V} = \mathbf{I}$.

5.2.7. Using the induced matrix norm (5.2.4), prove that if $\mathbf{A}$ is nonsingular, then

$$\|\mathbf{A}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \|\mathbf{A}^{-1}\mathbf{x}\|} \quad \text{or, equivalently,} \quad \|\mathbf{A}^{-1}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|}.$$

5.2.8. For $\mathbf{A} \in \mathcal{C}^{n\times n}$ and a parameter $z \in \mathcal{C}$, the matrix $\mathbf{R}(z) = (z\mathbf{I} - \mathbf{A})^{-1}$ is called the ***resolvent of*** $\mathbf{A}$. Prove that if $|z| > \|\mathbf{A}\|$ for any induced matrix norm, then

$$\|\mathbf{R}(z)\| \le \frac{1}{|z| - \|\mathbf{A}\|}.$$

5.3 INNER-PRODUCT SPACES

The euclidean norm, which naturally came first, is a coordinate-dependent concept. But by isolating its important properties we quickly moved to the more general coordinate-free definition of a vector norm given in (5.1.9) on p. 275. The goal is to now do the same for inner products. That is, start with the standard inner product, which is a coordinate-dependent definition, and identify properties that characterize the basic essence of the concept. The ones listed below are those that have been distilled from the standard inner product to formulate a more general coordinate-free definition.

General Inner Product

An ***inner product*** on a real (or complex) vector space $\mathcal{V}$ is a function that maps each ordered pair of vectors $\mathbf{x}, \mathbf{y}$ to a real (or complex) scalar $\langle \mathbf{x}|\mathbf{y} \rangle$ such that the following four properties hold.

$$\begin{aligned} &\langle \mathbf{x}|\mathbf{x} \rangle \text{ is real with } \langle \mathbf{x}|\mathbf{x} \rangle \geq 0, \text{ and } \langle \mathbf{x}|\mathbf{x} \rangle = 0 \text{ if and only if } \mathbf{x} = \mathbf{0}, \\ &\langle \mathbf{x}|\alpha\mathbf{y} \rangle = \alpha \langle \mathbf{x}|\mathbf{y} \rangle \text{ for all scalars } \alpha, \\ &\langle \mathbf{x}|\mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}|\mathbf{y} \rangle + \langle \mathbf{x}|\mathbf{z} \rangle, \\ &\langle \mathbf{x}|\mathbf{y} \rangle = \overline{\langle \mathbf{y}|\mathbf{x} \rangle} \quad \text{(for real spaces, this becomes } \langle \mathbf{x}|\mathbf{y} \rangle = \langle \mathbf{y}|\mathbf{x} \rangle). \end{aligned} \tag{5.3.1}$$

Notice that for each fixed value of $\mathbf{x}$, the second and third properties say that $\langle \mathbf{x}|\mathbf{y} \rangle$ is a linear function of $\mathbf{y}$.

Any real or complex vector space that is equipped with an inner product is called an ***inner-product space***.

Example 5.3.1

- The standard inner products, $\langle \mathbf{x}|\mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$ for $\Re^{n\times 1}$ and $\langle \mathbf{x}|\mathbf{y} \rangle = \mathbf{x}^*\mathbf{y}$ for $\mathcal{C}^{n\times 1}$, each satisfy the four defining conditions (5.3.1) for a general inner product—this shouldn't be a surprise.

- If $\mathbf{A}_{n\times n}$ is a nonsingular matrix, then $\langle \mathbf{x}|\mathbf{y} \rangle = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{y}$ is an inner product for $\mathcal{C}^{n\times 1}$. This inner product is sometimes called an $\mathbf{A}$***-inner product*** or an ***elliptical inner product.***

- Consider the vector space of $m \times n$ matrices. The functions defined by

$$\langle \mathbf{A}|\mathbf{B} \rangle = trace\left(\mathbf{A}^T\mathbf{B}\right) \quad \text{and} \quad \langle \mathbf{A}|\mathbf{B} \rangle = trace\left(\mathbf{A}^*\mathbf{B}\right) \tag{5.3.2}$$

  are inner products for $\Re^{m\times n}$ and $\mathcal{C}^{m\times n}$, respectively. These are referred to as the ***standard inner products for matrices***. Notice that these reduce to the standard inner products for vectors when $n = 1$.

- If $\mathcal{V}$ is the vector space of real-valued continuous functions defined on the interval (a, b), then
$$\langle f|g\rangle = \int_a^b f(t)g(t)dt$$
is an inner product on $\mathcal{V}$.

Just as the standard inner product for $\mathcal{C}^{n\times 1}$ defines the euclidean norm on $\mathcal{C}^{n\times 1}$ by $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^*\mathbf{x}}$, every general inner product in an inner-product space $\mathcal{V}$ defines a norm on $\mathcal{V}$ by setting

$$\|\star\| = \sqrt{\langle\star|\star\rangle}. \tag{5.3.3}$$

It's straightforward to verify that this satisfies the first two conditions in (5.2.3) on p. 280 that define a general vector norm, but, just as in the case of euclidean norms, verifying that (5.3.3) satisfies the triangle inequality requires a generalized version of CBS inequality.

General CBS Inequality

If $\mathcal{V}$ is an inner-product space, and if we set $\|\star\| = \sqrt{\langle\star|\star\rangle}$, then

$$|\langle\mathbf{x}|\mathbf{y}\rangle| \le \|\mathbf{x}\|\,\|\mathbf{y}\| \quad \text{for all } x, y \in \mathcal{V}. \tag{5.3.4}$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha = \langle\mathbf{x}|\mathbf{y}\rangle / \|\mathbf{x}\|^2$.

Proof. Set $\alpha = \langle\mathbf{x}|\mathbf{y}\rangle / \|\mathbf{x}\|^2$ (assume $\mathbf{x} \neq \mathbf{0}$, for otherwise there is nothing to prove), and observe that $\langle\mathbf{x}|\alpha\mathbf{x} - \mathbf{y}\rangle = 0$, so

$$\begin{aligned}
0 \le \|\alpha\mathbf{x} - \mathbf{y}\|^2 &= \langle\alpha\mathbf{x} - \mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle \\
&= \bar{\alpha}\,\langle\mathbf{x}|\alpha\mathbf{x} - \mathbf{y}\rangle - \langle\mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle \quad \text{(see Exercise 5.3.2)} \\
&= -\langle\mathbf{y}|\alpha\mathbf{x} - \mathbf{y}\rangle = \langle\mathbf{y}|\mathbf{y}\rangle - \alpha\,\langle\mathbf{y}|\mathbf{x}\rangle = \frac{\|\mathbf{y}\|^2\,\|\mathbf{x}\|^2 - \langle\mathbf{x}|\mathbf{y}\rangle\,\langle\mathbf{y}|\mathbf{x}\rangle}{\|\mathbf{x}\|^2}.
\end{aligned}$$

Since $\langle\mathbf{y}|\mathbf{x}\rangle = \overline{\langle\mathbf{x}|\mathbf{y}\rangle}$, it follows that $\langle\mathbf{x}|\mathbf{y}\rangle\,\langle\mathbf{y}|\mathbf{x}\rangle = |\langle\mathbf{x}|\mathbf{y}\rangle|^2$, so

$$0 \le \frac{\|\mathbf{y}\|^2\,\|\mathbf{x}\|^2 - |\langle\mathbf{x}|\mathbf{y}\rangle|^2}{\|\mathbf{x}\|^2} \implies |\langle\mathbf{x}|\mathbf{y}\rangle| \le \|\mathbf{x}\|\,\|\mathbf{y}\|.$$

Establishing the conditions for equality is the same as in Exercise 5.1.9. ∎

Let's now complete the job of showing that $\|\star\| = \sqrt{\langle\star|\star\rangle}$ is indeed a vector norm as defined in (5.2.3) on p. 280.

Norms in Inner-Product Spaces

If $\mathcal{V}$ is an inner-product space with an inner product $\langle \mathbf{x}|\mathbf{y}\rangle$, then

$$\|\star\| = \sqrt{\langle \star|\star\rangle} \quad \text{defines a norm on } \mathcal{V}.$$

Proof. The fact that $\|\star\| = \sqrt{\langle \star|\star\rangle}$ satisfies the first two norm properties in (5.2.3) on p. 280 follows directly from the defining properties (5.3.1) for an inner product. You are asked to provide the details in Exercise 5.3.3. To establish the triangle inequality, use $\langle \mathbf{x}|\mathbf{y}\rangle \le |\langle \mathbf{x}|\mathbf{y}\rangle|$ and $\langle \mathbf{y}|\mathbf{x}\rangle = \overline{\langle \mathbf{x}|\mathbf{y}\rangle} \le |\langle \mathbf{x}|\mathbf{y}\rangle|$ together with the CBS inequality to write

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{x}\rangle + \langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{y}|\mathbf{x}\rangle + \langle \mathbf{y}|\mathbf{y}\rangle \\ &\le \|\mathbf{x}\|^2 + 2|\langle \mathbf{x}|\mathbf{y}\rangle| + \|\mathbf{y}\|^2 \le (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \quad \blacksquare\end{aligned}$$

Example 5.3.2

Problem: Describe the norms that are generated by the inner products presented in Example 5.3.1.

- Given a nonsingular matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$, the $\mathbf{A}$-***norm*** (or ***elliptical norm***) generated by the $\mathbf{A}$-inner product on $\mathcal{C}^{n\times 1}$ is

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle} = \sqrt{\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}} = \|\mathbf{A}\mathbf{x}\|_2 . \tag{5.3.5}$$

- The standard inner product for matrices generates the Frobenius matrix norm because

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}|\mathbf{A}\rangle} = \sqrt{trace\,(\mathbf{A}^*\mathbf{A})} = \|\mathbf{A}\|_F . \tag{5.3.6}$$

- For the space of real-valued continuous functions defined on (a,b), the norm of a function f generated by the inner product $\langle f|g\rangle = \int_a^b f(t)g(t)dt$ is

$$\|f\| = \sqrt{\langle f|f\rangle} = \left(\int_a^b f(t)^2 dt\right)^{1/2}.$$

Example 5.3.3

To illustrate the utility of the ideas presented above, consider the proposition

$$trace\left(\mathbf{A}^T\mathbf{B}\right)^2 \leq trace\left(\mathbf{A}^T\mathbf{A}\right) trace\left(\mathbf{B}^T\mathbf{B}\right) \quad \text{for all } \mathbf{A}, \mathbf{B} \in \Re^{m\times n}.$$

Problem: How would you know to formulate such a proposition and, second, how do you prove it?

Solution: The answer to both questions is the same. This is the CBS inequality in $\Re^{m\times n}$ equipped with the standard inner product $\langle \mathbf{A}|\mathbf{B}\rangle = trace\left(\mathbf{A}^T\mathbf{B}\right)$ and associated norm $\|\mathbf{A}\|_F = \sqrt{\langle \mathbf{A}|\mathbf{A}\rangle} = \sqrt{trace\left(\mathbf{A}^T\mathbf{A}\right)}$ because CBS says

$$\langle \mathbf{A}|\mathbf{B}\rangle^2 \leq \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 \quad \Longrightarrow \quad trace\left(\mathbf{A}^T\mathbf{B}\right)^2 \leq trace\left(\mathbf{A}^T\mathbf{A}\right) trace\left(\mathbf{B}^T\mathbf{B}\right).$$

The point here is that if your knowledge is limited to elementary matrix manipulations (which is all that is needed to understand the statement of the proposition), formulating the correct inequality might be quite a challenge to your intuition. And then proving the proposition using only elementary matrix manipulations would be a significant task—essentially, you would have to derive a version of CBS. But knowing the basic facts of inner-product spaces makes the proposition nearly trivial to conjecture and prove.

Since each inner product generates a norm by the rule $\|\star\| = \sqrt{\langle \star|\star\rangle}$, it's natural to ask if the reverse is also true. That is, for each vector norm $\|\star\|$ on a space $\mathcal{V}$, does there exist a corresponding inner product on $\mathcal{V}$ such that $\sqrt{\langle \star|\star\rangle} = \|\star\|$? If not, under what conditions will a given norm be generated by an inner product? These are tricky questions, and it took the combined efforts of Maurice R. Fréchet[38] (1878–1973) and John von Neumann (1903–1957) to provide the answer.

[38] Maurice René Fréchet began his illustrious career by writing an outstanding Ph.D. dissertation in 1906 under the direction of the famous French mathematician Jacques Hadamard (p. 469) in which the concepts of a metric space and compactness were first formulated. Fréchet developed into a versatile mathematical scientist, and he served as professor of mechanics at the University of Poitiers (1910–1919), professor of higher calculus at the University of Strasbourg (1920–1927), and professor of differential and integral calculus and professor of the calculus of probabilities at the University of Paris (1928–1948).

Born in Budapest, Hungary, John von Neumann was a child prodigy who could divide eight-digit numbers in his head when he was only six years old. Due to the political unrest in Europe, he came to America, where, in 1933, he became one of the six original professors of mathematics at the Institute for Advanced Study at Princeton University, a position he retained for the rest of his life. During his career, von Neumann's genius touched mathematics (pure and applied), chemistry, physics, economics, and computer science, and he is generally considered to be among the best scientists and mathematicians of the twentieth century.

Parallelogram Identity

For a given norm $\|\star\|$ on a vector space $\mathcal{V}$, there exists an inner product on $\mathcal{V}$ such that $\langle\star|\star\rangle = \|\star\|^2$ if and only if the ***parallelogram identity***

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\big(\,\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\,\big) \tag{5.3.7}$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

Proof. Consider real spaces—complex spaces are discussed in Exercise 5.3.6. If there exists an inner product such that $\langle\star|\star\rangle = \|\star\|^2$, then the parallelogram identity is immediate because $\langle\mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle + \langle\mathbf{x}-\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle = 2\,\langle\mathbf{x}|\mathbf{x}\rangle + 2\,\langle\mathbf{y}|\mathbf{y}\rangle$. The difficult part is establishing the converse. Suppose $\|\star\|$ satisfies the parallelogram identity, and prove that the function

$$\langle\mathbf{x}|\mathbf{y}\rangle = \frac{1}{4}\big(\,\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\,\big) \tag{5.3.8}$$

is an inner product for $\mathcal{V}$ such that $\langle\mathbf{x}|\mathbf{x}\rangle = \|\mathbf{x}\|^2$ for all $\mathbf{x}$ by showing the four defining conditions (5.3.1) hold. The first and fourth conditions are immediate. To establish the third, use the parallelogram identity to write

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 &= \frac{1}{2}\big(\,\|\mathbf{x}+\mathbf{y}+\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2\,\big),\\
\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 &= \frac{1}{2}\big(\,\|\mathbf{x}-\mathbf{y}+\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{z}-\mathbf{y}\|^2\,\big),
\end{aligned}$$

and then subtract to obtain

$$\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 = \frac{\|2\mathbf{x}+(\mathbf{y}+\mathbf{z})\|^2 - \|2\mathbf{x}-(\mathbf{y}+\mathbf{z})\|^2}{2}.$$

Consequently,

$$\begin{aligned}
\langle\mathbf{x}|\mathbf{y}\rangle + \langle\mathbf{x}|\mathbf{z}\rangle &= \frac{1}{4}\big(\,\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2\,\big)\\
&= \frac{1}{8}\big(\,\|2\mathbf{x}+(\mathbf{y}+\mathbf{z})\|^2 - \|2\mathbf{x}-(\mathbf{y}+\mathbf{z})\|^2\,\big)\\
&= \frac{1}{2}\left(\left\|\mathbf{x}+\frac{\mathbf{y}+\mathbf{z}}{2}\right\|^2 - \left\|\mathbf{x}-\frac{\mathbf{y}+\mathbf{z}}{2}\right\|^2\right) = 2\left\langle\mathbf{x}\,\middle|\,\frac{\mathbf{y}+\mathbf{z}}{2}\right\rangle,
\end{aligned} \tag{5.3.9}$$

and setting $\mathbf{z} = \mathbf{0}$ produces the statement that $\langle\mathbf{x}|\mathbf{y}\rangle = 2\,\langle\mathbf{x}|\mathbf{y}/2\rangle$ for all $\mathbf{y} \in \mathcal{V}$. Replacing $\mathbf{y}$ by $\mathbf{y}+\mathbf{z}$ yields $\langle\mathbf{x}|\mathbf{y}+\mathbf{z}\rangle = 2\,\langle\mathbf{x}|(\mathbf{y}+\mathbf{z})/2\rangle$, and thus (5.3.9)

guarantees that $\langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{x}|\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{y}+\mathbf{z}\rangle$. Now prove that $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle$ for all real α. This is valid for integer values of α by the result just established, and it holds when α is rational because if β and γ are integers, then

$$\gamma^2 \left\langle \mathbf{x} \middle| \frac{\beta}{\gamma}\mathbf{y} \right\rangle = \langle \gamma\mathbf{x}|\beta\mathbf{y}\rangle = \beta\gamma\langle \mathbf{x}|\mathbf{y}\rangle \quad \Longrightarrow \quad \left\langle \mathbf{x} \middle| \frac{\beta}{\gamma}\mathbf{y} \right\rangle = \frac{\beta}{\gamma}\langle \mathbf{x}|\mathbf{y}\rangle .$$

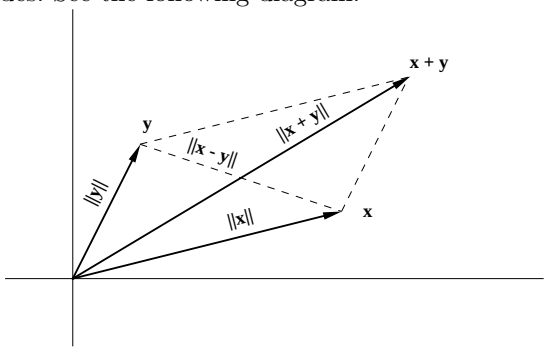
Because $\|\mathbf{x}+\alpha\mathbf{y}\|$ and $\|\mathbf{x}-\alpha\mathbf{y}\|$ are continuous functions of α (Exercise 5.1.7), equation (5.3.8) insures that $\langle \mathbf{x}|\alpha\mathbf{y}\rangle$ is a continuous function of α. Therefore, if α is irrational, and if $\{\alpha_n\}$ is a sequence of rational numbers such that $\alpha_n \to \alpha$, then $\langle \mathbf{x}|\alpha_n\mathbf{y}\rangle \to \langle \mathbf{x}|\alpha\mathbf{y}\rangle$ and $\langle \mathbf{x}|\alpha_n\mathbf{y}\rangle = \alpha_n\langle \mathbf{x}|\mathbf{y}\rangle \to \alpha\langle \mathbf{x}|\mathbf{y}\rangle$, so $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle$. ∎

Example 5.3.4

We already know that the euclidean vector norm on $\mathcal{C}^n$ is generated by the standard inner product, so the previous theorem guarantees that the parallelogram identity must hold for the 2-norm. This is easily corroborated by observing that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2^2 + \|\mathbf{x}-\mathbf{y}\|_2^2 &= (\mathbf{x}+\mathbf{y})^*(\mathbf{x}+\mathbf{y}) + (\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y}) \\ &= 2\,(\mathbf{x}^*\mathbf{x}+\mathbf{y}^*\mathbf{y}) = 2(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2).\end{aligned}$$

The parallelogram identity is so named because it expresses the fact that the sum of the squares of the diagonals in a parallelogram is twice the sum of the squares of the sides. See the following diagram.



Example 5.3.5

Problem: Except for the euclidean norm, is any other vector p-norm generated by an inner product?

Solution: No, because the parallelogram identity (5.3.7) doesn't hold when $p \neq 2$. To see that $\|\mathbf{x}+\mathbf{y}\|_p^2 + \|\mathbf{x}-\mathbf{y}\|_p^2 = 2\big(\|\mathbf{x}\|_p^2 + \|\mathbf{y}\|_p^2\big)$ is not valid for all $\mathbf{x},\mathbf{y} \in \mathcal{C}^n$ when $p \neq 2$, consider $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = \mathbf{e}_2$. It's apparent that $\|\mathbf{e_1}+\mathbf{e_2}\|_p^2 = 2^{2/p} = \|\mathbf{e_1}-\mathbf{e_2}\|_p^2$, so

$$\|\mathbf{e_1}+\mathbf{e_2}\|_p^2 + \|\mathbf{e_1}-\mathbf{e_2}\|_p^2 = 2^{(p+2)/p} \quad \text{and} \quad 2\big(\|\mathbf{e_1}\|_p^2 + \|\mathbf{e_2}\|_p^2\big) = 4.$$

Clearly, $2^{(p+2)/p} = 4$ only when $p = 2$. Details for the ∞-norm are asked for in Exercise 5.3.7.

Conclusion: For applications that are best analyzed in the context of an inner-product space (e.g., least squares problems), we are limited to the euclidean norm or else to one of its variation such as the elliptical norm in (5.3.5).

Virtually all important statements concerning $\Re^n$ or $\mathcal{C}^n$ with the standard inner product remain valid for general inner-product spaces—e.g., consider the statement and proof of the general CBS inequality. Advanced or more theoretical texts prefer a development in terms of general inner-product spaces. However, the focus of this text is matrices and the coordinate spaces $\Re^n$ and $\mathcal{C}^n$, so subsequent discussions will usually be phrased in terms of $\Re^n$ or $\mathcal{C}^n$ and their standard inner products. But remember that extensions to more general inner-product spaces are always lurking in the background, and we will not hesitate to use these generalities or general inner-product notation when they serve our purpose.

Exercises for section 5.3

5.3.1. For $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$, determine which of the following are inner products for $\Re^{3\times 1}$.

(a) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1y_1 + x_3y_3$,
(b) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1y_1 - x_2y_2 + x_3y_3$,
(c) $\langle \mathbf{x}|\mathbf{y}\rangle = 2x_1y_1 + x_2y_2 + 4x_3y_3$,
(d) $\langle \mathbf{x}|\mathbf{y}\rangle = x_1^2y_1^2 + x_2^2y_2^2 + x_3^2y_3^2$.

5.3.2. For a general inner-product space $\mathcal{V}$, explain why each of the following statements must be true.

(a) If $\langle \mathbf{x}|\mathbf{y}\rangle = 0$ for all $\mathbf{x} \in \mathcal{V}$, then $\mathbf{y} = \mathbf{0}$.
(b) $\langle \alpha\mathbf{x}|\mathbf{y}\rangle = \overline{\alpha}\,\langle \mathbf{x}|\mathbf{y}\rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and for all scalars α.
(c) $\langle \mathbf{x}+\mathbf{y}|\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{z}\rangle + \langle \mathbf{y}|\mathbf{z}\rangle$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$.

5.3.3. Let $\mathcal{V}$ be an inner-product space with an inner product $\langle \mathbf{x}|\mathbf{y}\rangle$. Explain why the function defined by $\|\star\| = \sqrt{\langle \star|\star\rangle}$ satisfies the first two norm properties in (5.2.3) on p. 280.

5.3.4. For a real inner-product space with $\|\star\|^2 = \langle \star|\star\rangle$, derive the inequality

$$\langle \mathbf{x}|\mathbf{y}\rangle \le \frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{2}. \qquad \textbf{Hint: } \text{Consider } \mathbf{x} - \mathbf{y}.$$

5.3.5. For $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, explain why each of the following inequalities is valid.

(a) $|trace\,(\mathbf{B})|^2 \leq n\,[trace\,(\mathbf{B}^*\mathbf{B})]\,.$

(b) $trace\left(\mathbf{B}^2\right) \leq trace\left(\mathbf{B}^T\mathbf{B}\right)$ for real matrices.

(c) $trace\left(\mathbf{A}^T\mathbf{B}\right) \leq \dfrac{trace\left(\mathbf{A}^T\mathbf{A}\right) + trace\left(\mathbf{B}^T\mathbf{B}\right)}{2}$ for real matrices.

5.3.6. Extend the proof given on p. 290 concerning the parallelogram identity (5.3.7) to include complex spaces. **Hint:** If $\mathcal{V}$ is a complex space with a norm $\|\star\|$ that satisfies the parallelogram identity, let

$$\langle \mathbf{x}|\mathbf{y}\rangle_r = \frac{\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2}{4},$$

and prove that

$$\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{x}|\mathbf{y}\rangle_r + \mathrm{i}\,\langle \mathrm{i}\mathbf{x}|\mathbf{y}\rangle_r \quad \text{(the } \textbf{\textit{polarization identity}}\text{)} \tag{5.3.10}$$

is an inner product on $\mathcal{V}$.

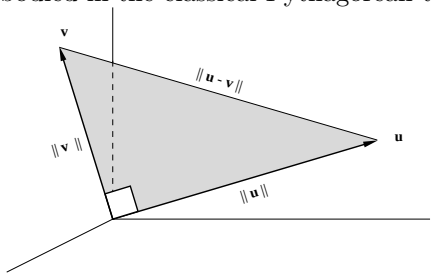
5.3.7. Explain why there does not exist an inner product on $\mathcal{C}^n$ $(n \geq 2)$ such that $\|\star\|_\infty = \sqrt{\langle \star|\star\rangle}$.

5.3.8. Explain why the Frobenius matrix norm on $\mathcal{C}^{n\times n}$ must satisfy the parallelogram identity.

5.3.9. For $n \geq 2$, is either the matrix 1-, 2-, or ∞-norm generated by an inner product on $\mathcal{C}^{n\times n}$?

5.4 ORTHOGONAL VECTORS

Two vectors in $\Re^3$ are ***orthogonal*** (perpendicular) if the angle between them is a right angle (90°). But the visual concept of a right angle is not at our disposal in higher dimensions, so we must dig a little deeper. The essence of perpendicularity in $\Re^2$ and $\Re^3$ is embodied in the classical Pythagorean theorem,



which says that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \|\mathbf{u}-\mathbf{v}\|^2$. But [39] $\|\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{u}$ for all $\mathbf{u} \in \Re^3$, and $\mathbf{u}^T\mathbf{v} = \mathbf{v}^T\mathbf{u}$, so we can rewrite the Pythagorean statement as

$$\begin{aligned}\mathbf{0} &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2 = \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v}) \\ &= \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - \left(\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{v} - \mathbf{v}^T\mathbf{u} + \mathbf{v}^T\mathbf{v}\right) = 2\mathbf{u}^T\mathbf{v}.\end{aligned}$$

Therefore, $\mathbf{u}$ and $\mathbf{v}$ are orthogonal vectors in $\Re^3$ if and only if $\mathbf{u}^T\mathbf{v} = 0$. The natural extension of this provides us with a definition in more general spaces.

Orthogonality

In an inner-product space $\mathcal{V}$, two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ are said to be ***orthogonal*** (to each other) whenever $\langle \mathbf{x} | \mathbf{y} \rangle = 0$, and this is denoted by writing $\mathbf{x} \perp \mathbf{y}$.

- For $\Re^n$ with the standard inner product, $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^T\mathbf{y} = 0$.
- For $\mathcal{C}^n$ with the standard inner product, $\mathbf{x} \perp \mathbf{y} \iff \mathbf{x}^*\mathbf{y} = 0$.

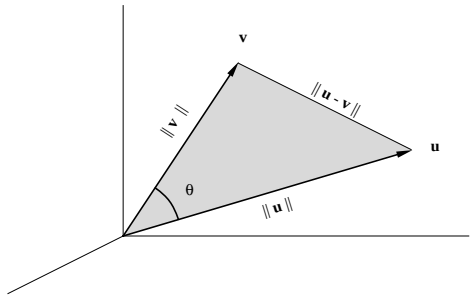
Example 5.4.1

$$\mathbf{x} = \begin{pmatrix} 1 \\ -2 \\ 3 \\ -1 \end{pmatrix} \text{ is orthogonal to } \mathbf{y} = \begin{pmatrix} 4 \\ 1 \\ -2 \\ -4 \end{pmatrix} \text{ because } \mathbf{x}^T\mathbf{y} = 0.$$

[39] Throughout this section, only norms generated by an underlying inner product $\|\star\|^2 = \langle \star | \star \rangle$ are used, so distinguishing subscripts on the norm notation can be omitted.

In spite of the fact that $\mathbf{u}^T\mathbf{v} = 0$, the vectors $\mathbf{u} = \begin{pmatrix} \mathrm{i} \\ 3 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} \mathrm{i} \\ 0 \\ 1 \end{pmatrix}$ are *not* orthogonal because $\mathbf{u}^*\mathbf{v} \neq 0$.

Now that "right angles" in higher dimensions make sense, how can more general angles be defined? Proceed just as before, but use the law of cosines rather than the Pythagorean theorem. Recall that



the ***law of cosines*** in $\Re^2$ or $\Re^3$ says $\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\,\|\mathbf{u}\|\,\|\mathbf{v}\|\,\cos\theta$. If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, then this reduces to the Pythagorean theorem. But, in general,

$$\cos\theta = \frac{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2}{2\,\|\mathbf{u}\|\,\|\mathbf{v}\|} = \frac{\mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} - (\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})}{2\,\|\mathbf{u}\|\,\|\mathbf{v}\|}$$
$$= \frac{2\mathbf{u}^T\mathbf{v}}{2\,\|\mathbf{u}\|\,\|\mathbf{v}\|} = \frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

This easily extends to higher dimensions because if $\mathbf{x}$, $\mathbf{y}$ are vectors from any real inner-product space, then the general CBS inequality (5.3.4) on p. 287 guarantees that $\langle \mathbf{x}|\mathbf{y}\rangle \,/\, \|\mathbf{x}\|\,\|\mathbf{y}\|$ is a number in the interval $[-1,\,1]$, and hence there is a unique value θ in $[0,\,\pi]$ such that $\cos\theta = \langle \mathbf{x}|\mathbf{y}\rangle \,/\, \|\mathbf{x}\|\,\|\mathbf{y}\|$.

Angles

In a real inner-product space $\mathcal{V}$, the radian measure of the ***angle*** between nonzero vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ is defined to be the number $\theta \in [0,\,\pi]$ such that

$$\cos\theta = \frac{\langle \mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|\,\|\mathbf{y}\|}. \tag{5.4.1}$$

Example 5.4.2

In $\Re^n$, $\cos\theta = \mathbf{x}^T\mathbf{y}/\,\|\mathbf{x}\|\,\|\mathbf{y}\|$. For example, to determine the angle between $\mathbf{x} = \begin{pmatrix} -4 \\ 2 \\ 1 \\ 2 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \end{pmatrix}$, compute $\cos\theta = 2/(5)(3) = 2/15$, and use the inverse cosine function to conclude that $\theta = 1.437$ radians (rounded).

Example 5.4.3

Linear Correlation. Suppose that an experiment is conducted, and the resulting observations are recorded in two data vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{and let } \mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Problem: Determine to what extent the y_i's are linearly related to the x_i's. That is, measure how close $\mathbf{y}$ is to being a linear combination $\beta_0\mathbf{e} + \beta_1\mathbf{x}$.

Solution: The cosine as defined in (5.4.1) does the job. To understand how, let $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ be the ***mean*** and ***standard deviation*** of the data in $\mathbf{x}$. That is,

$$\mu_{\mathbf{x}} = \frac{\sum_i x_i}{n} = \frac{\mathbf{e}^T\mathbf{x}}{n} \quad \text{and} \quad \sigma_{\mathbf{x}} = \sqrt{\frac{\sum_i (x_i - \mu_{\mathbf{x}})^2}{n}} = \frac{\|\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}\|_2}{\sqrt{n}}.$$

The mean is a measure of central tendency, and the standard deviation measures the extent to which the data is spread. Frequently, raw data from different sources is difficult to compare because the units of measure are different—e.g., one researcher may use the metric system while another uses American units. To compensate, data is almost always first "standardized" into unitless quantities. The ***standardization*** of a vector $\mathbf{x}$ for which $\sigma_{\mathbf{x}} \neq 0$ is defined to be

$$\mathbf{z}_{\mathbf{x}} = \frac{\mathbf{x} - \mu_{\mathbf{x}}\mathbf{e}}{\sigma_{\mathbf{x}}}.$$

Entries in $\mathbf{z}_{\mathbf{x}}$ are often referred to as ***standard scores*** or ***z-scores.*** All standardized vectors have the properties that $\|\mathbf{z}\| = \sqrt{n}$, $\mu_{\mathbf{z}} = 0$, and $\sigma_{\mathbf{z}} = 1$. Furthermore, it's not difficult to verify that for vectors $\mathbf{x}$ and $\mathbf{y}$ such that $\sigma_{\mathbf{x}} \neq 0$ and $\sigma_{\mathbf{y}} \neq 0$, it's the case that

$$\begin{aligned} \mathbf{z}_{\mathbf{x}} = \mathbf{z}_{\mathbf{y}} &\iff \exists \text{ constants } \beta_0,\, \beta_1 \text{ such that } \mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}, \quad \text{where} \quad \beta_1 > 0, \\ \mathbf{z}_{\mathbf{x}} = -\mathbf{z}_{\mathbf{y}} &\iff \exists \text{ constants } \beta_0,\, \beta_1 \text{ such that } \mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}, \quad \text{where} \quad \beta_1 < 0. \end{aligned}$$

- In other words, $\mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}$ for some β_0 and β_1 if and only if $\mathbf{z}_{\mathbf{x}} = \pm\mathbf{z}_{\mathbf{y}}$, in which case we say $\mathbf{y}$ is ***perfectly linearly correlated*** with $\mathbf{x}$.

Since $\mathbf{z_x}$ varies continuously with $\mathbf{x}$, the existence of a "near" linear relationship between $\mathbf{x}$ and $\mathbf{y}$ is equivalent to $\mathbf{z_x}$ being "close" to $\pm\mathbf{z_y}$ in some sense. The fact that $\|\mathbf{z_x}\| = \|\pm\mathbf{z_y}\| = \sqrt{n}$ means $\mathbf{z_x}$ and $\pm\mathbf{z_y}$ differ only in orientation, so a natural measure of how close $\mathbf{z_x}$ is to $\pm\mathbf{z_y}$ is $\cos\theta$, where θ is the angle between $\mathbf{z_x}$ and $\mathbf{z_y}$. The number

$$\rho_{\mathbf{xy}} = \cos\theta = \frac{\mathbf{z_x}^T\mathbf{z_y}}{\|\mathbf{z_x}\|\ \|\mathbf{z_y}\|} = \frac{\mathbf{z_x}^T\mathbf{z_y}}{n} = \frac{(\mathbf{x}-\mu_{\mathbf{x}}\mathbf{e})^T(\mathbf{y}-\mu_{\mathbf{y}}\mathbf{e})}{\|\mathbf{x}-\mu_{\mathbf{x}}\mathbf{e}\|\ \|\mathbf{y}-\mu_{\mathbf{y}}\mathbf{e}\|}$$

is called the ***coefficient of linear correlation,*** and the following facts are now immediate.

- $\rho_{\mathbf{xy}} = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, in which case we say that $\mathbf{x}$ and $\mathbf{y}$ are ***completely uncorrelated.***
- $|\rho_{\mathbf{xy}}| = 1$ if and only if $\mathbf{y}$ is *perfectly* correlated with $\mathbf{x}$. That is, $|\rho_{\mathbf{xy}}| = 1$ if and only if there exists a linear relationship $\mathbf{y} = \beta_0\mathbf{e} + \beta_1\mathbf{x}$.
  - ▷ When $\beta_1 > 0$, we say that $\mathbf{y}$ is ***positively correlated*** with $\mathbf{x}$.
  - ▷ When $\beta_1 < 0$, we say that $\mathbf{y}$ is ***negatively correlated*** with $\mathbf{x}$.
- $|\rho_{\mathbf{xy}}|$ measures the degree to which $\mathbf{y}$ is linearly related to $\mathbf{x}$. In other words, $|\rho_{\mathbf{xy}}| \approx 1$ if and only if $\mathbf{y} \approx \beta_0\mathbf{e} + \beta_1\mathbf{x}$ for some β_0 and β_1.
  - ▷ Positive correlation is measured by the degree to which $\rho_{\mathbf{xy}} \approx 1$.
  - ▷ Negative correlation is measured by the degree to which $\rho_{\mathbf{xy}} \approx -1$.

If the data in $\mathbf{x}$ and $\mathbf{y}$ are plotted in $\Re^2$ as points (x_i, y_i), then, as depicted in Figure 5.4.1, $\rho_{\mathbf{xy}} \approx 1$ means that the points lie near a straight line with positive slope, while $\rho_{\mathbf{xy}} \approx -1$ means that the points lie near a line with negative slope, and $\rho_{\mathbf{xy}} \approx 0$ means that the points do not lie near a straight line.

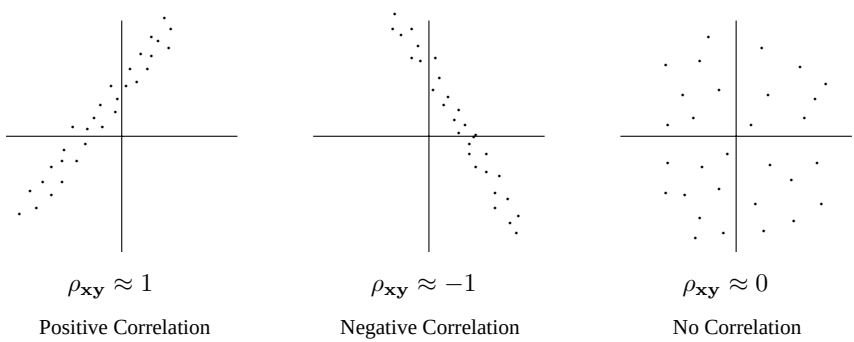


FIGURE 5.4.1

If $|\rho_{\mathbf{xy}}| \approx 1$, then the theory of least squares as presented in §4.6 can be used to determine a "best-fitting" straight line.

Orthonormal Sets

$\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is called an ***orthonormal set*** whenever $\|\mathbf{u}_i\| = 1$ for each i, and $\mathbf{u}_i \perp \mathbf{u}_j$ for all $i \neq j$. In other words,

$$\langle \mathbf{u}_i | \mathbf{u}_j \rangle = \begin{cases} 1 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

- Every orthonormal set is linearly independent. (5.4.2)
- Every orthonormal set of n vectors from an n-dimensional space $\mathcal{V}$ is an orthonormal basis for $\mathcal{V}$.

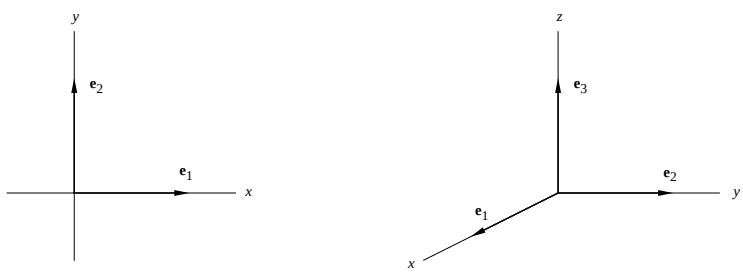
Proof. The second point follows from the first. To prove the first statement, suppose $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is orthonormal. If $\mathbf{0} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n$, use the properties of an inner product to write

$$\begin{aligned} 0 = \langle \mathbf{u}_i | \mathbf{0} \rangle &= \langle \mathbf{u}_i | \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n \rangle \\ &= \alpha_1 \langle \mathbf{u}_i | \mathbf{u}_1 \rangle + \cdots + \alpha_i \langle \mathbf{u}_i | \mathbf{u}_i \rangle + \cdots + \alpha_n \langle \mathbf{u}_i | \mathbf{u}_n \rangle = \alpha_i \|\mathbf{u}_i\|^2 \\ &= \alpha_i \quad \text{for each } i. \quad \blacksquare \end{aligned}$$

Example 5.4.4

The set $\mathcal{B}' = \left\{ \mathbf{u}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}$ is a set of mutually orthogonal vectors because $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$, but $\mathcal{B}'$ is *not* an orthonormal set—each vector does not have unit length. However, it's easy to convert an orthogonal set (not containing a zero vector) into an orthonormal set by simply normalizing each vector. Since $\|\mathbf{u}_1\| = \sqrt{2}$, $\|\mathbf{u}_2\| = \sqrt{3}$, and $\|\mathbf{u}_3\| = \sqrt{6}$, it follows that $\mathcal{B} = \left\{ \mathbf{u}_1/\sqrt{2}, \mathbf{u}_2/\sqrt{3}, \mathbf{u}_3/\sqrt{6} \right\}$ is orthonormal.

The most common orthonormal basis is $\mathcal{S} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, the standard basis for $\Re^n$ and $\mathcal{C}^n$, and, as illustrated below for $\Re^2$ and $\Re^3$, these orthonormal vectors are directed along the standard coordinate axes.



Another orthonormal basis $\mathcal{B}$ need not be directed in the same way as $\mathcal{S}$, but that's the only significant difference because it's geometrically evident that $\mathcal{B}$ must amount to some rotation of $\mathcal{S}$. Consequently, we should expect general orthonormal bases to provide essentially the same advantages as the standard basis. For example, an important function of the standard basis $\mathcal{S}$ for $\Re^n$ is to provide coordinate representations by writing

$$\mathbf{x} = [\mathbf{x}]_{\mathcal{S}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{to mean} \quad \mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

With respect to a general basis $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$, the coordinates of $\mathbf{x}$ are the scalars ξ_i in the representation $\mathbf{x} = \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \cdots + \xi_n\mathbf{u}_n$, and, as illustrated in Example 4.7.2, finding the ξ_i's requires solving an $n \times n$ system, a nuisance we would like to avoid. But if $\mathcal{B}$ is an *orthonormal* basis, then the ξ_i's are readily available because $\langle \mathbf{u}_i | \mathbf{x} \rangle = \langle \mathbf{u}_i | \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \cdots + \xi_n\mathbf{u}_n \rangle = \sum_{j=1}^{n} \xi_j \langle \mathbf{u}_i | \mathbf{u}_j \rangle = \xi_i \|\mathbf{u}_i\|^2 = \xi_i$. This yields the ***Fourier***[40] ***expansion*** of $\mathbf{x}$.

Fourier Expansions

If $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is an orthonormal basis for an inner-product space $\mathcal{V}$, then each $\mathbf{x} \in \mathcal{V}$ can be expressed as

$$\mathbf{x} = \langle \mathbf{u}_1 | \mathbf{x} \rangle \mathbf{u}_1 + \langle \mathbf{u}_2 | \mathbf{x} \rangle \mathbf{u}_2 + \cdots + \langle \mathbf{u}_n | \mathbf{x} \rangle \mathbf{u}_n. \tag{5.4.3}$$

This is called the ***Fourier expansion*** of $\mathbf{x}$. The scalars $\xi_i = \langle \mathbf{u}_i | \mathbf{x} \rangle$ are the coordinates of $\mathbf{x}$ with respect to $\mathcal{B}$, and they are called the ***Fourier coefficients.*** Geometrically, the Fourier expansion resolves $\mathbf{x}$ into n mutually orthogonal vectors $\langle \mathbf{u}_i | \mathbf{x} \rangle \mathbf{u}_i$, each of which represents the orthogonal projection of $\mathbf{x}$ onto the space (line) spanned by $\mathbf{u}_i$. (More is said in Example 5.13.1 on p. 431 and Exercise 5.13.11.)

[40] Jean Baptiste Joseph Fourier (1768–1830) was a French mathematician and physicist who, while studying heat flow, developed expansions similar to (5.4.3). Fourier's work dealt with special infinite-dimensional inner-product spaces involving trigonometric functions as discussed in Example 5.4.6. Although they were apparently used earlier by Daniel Bernoulli (1700–1782) to solve problems concerned with vibrating strings, these orthogonal expansions became known as ***Fourier series,*** and they are now a fundamental tool in applied mathematics. Born the son of a tailor, Fourier was orphaned at the age of eight. Although he showed a great aptitude for mathematics at an early age, he was denied his dream of entering the French artillery because of his "low birth." Instead, he trained for the priesthood, but he never took his vows. However, his talents did not go unrecognized, and he later became a favorite of Napoleon. Fourier's work is now considered as marking an epoch in the history of both pure and applied mathematics. The next time you are in Paris, check out Fourier's plaque on the first level of the Eiffel Tower.

Example 5.4.5

Problem: Determine the Fourier expansion of $\mathbf{x} = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$ with respect to the standard inner product and the orthonormal basis given in Example 5.4.4

$$\mathcal{B} = \left\{ \mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \mathbf{u}_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}.$$

Solution: The Fourier coefficients are

$$\xi_1 = \langle \mathbf{u}_1 | \mathbf{x} \rangle = \frac{-3}{\sqrt{2}}, \quad \xi_2 = \langle \mathbf{u}_2 | \mathbf{x} \rangle = \frac{2}{\sqrt{3}}, \quad \xi_3 = \langle \mathbf{u}_3 | \mathbf{x} \rangle = \frac{1}{\sqrt{6}},$$

so

$$\mathbf{x} = \xi_1 \mathbf{u}_1 + \xi_2 \mathbf{u}_2 + \xi_3 \mathbf{u}_3 = \frac{1}{2} \begin{pmatrix} -3 \\ 3 \\ 0 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + \frac{1}{6} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

You may find it instructive to sketch a picture of these vectors in $\Re^3$.

Example 5.4.6

Fourier Series. Let $\mathcal{V}$ be the inner-product space of real-valued functions that are integrable on the interval $(-\pi, \pi)$ and where the inner product and norm are given by

$$\langle f | g \rangle = \int_{-\pi}^{\pi} f(t) g(t) dt \quad \text{and} \quad \|f\| = \left(\int_{-\pi}^{\pi} f^2(t) dt \right)^{1/2}.$$

It's straightforward to verify that the set of trigonometric functions

$$\mathcal{B}' = \{1, \cos t, \cos 2t, \ldots, \sin t, \sin 2t, \sin 3t, \ldots\}$$

is a set of mutually orthogonal vectors, so normalizing each vector produces the orthonormal set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{\cos t}{\sqrt{\pi}}, \frac{\cos 2t}{\sqrt{\pi}}, \ldots, \frac{\sin t}{\sqrt{\pi}}, \frac{\sin 2t}{\sqrt{\pi}}, \frac{\sin 3t}{\sqrt{\pi}}, \ldots \right\}.$$

Given an arbitrary $f \in \mathcal{V}$, we construct its Fourier expansion

$$F(t) = \alpha_0 \frac{1}{\sqrt{2\pi}} + \sum_{k=1}^{\infty} \alpha_k \frac{\cos kt}{\sqrt{\pi}} + \sum_{k=1}^{\infty} \beta_k \frac{\sin kt}{\sqrt{\pi}}, \tag{5.4.4}$$

where the Fourier coefficients are given by

$$\alpha_0 = \left\langle \frac{1}{\sqrt{2\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(t)dt\,,$$
$$\alpha_k = \left\langle \frac{\cos kt}{\sqrt{\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(t)\cos kt\,dt \quad \text{for } k = 1, 2, 3, \ldots,$$
$$\beta_k = \left\langle \frac{\sin kt}{\sqrt{\pi}} \middle| f \right\rangle = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} f(t)\sin kt\,dt \quad \text{for } k = 1, 2, 3, \ldots.$$

Substituting these coefficients in (5.4.4) produces the infinite series

$$F(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nt + b_n \sin nt)\,, \tag{5.4.5}$$

where

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\cos nt\,dt \quad \text{and} \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)\sin nt\,dt. \tag{5.4.6}$$

The series $F(t)$ in (5.4.5) is called the ***Fourier series*** expansion for $f(t)$, but, unlike the situation in finite-dimensional spaces, $F(t)$ need not agree with the original function $f(t)$. After all, F is periodic, so there is no hope of agreement when f is not periodic. However, the following statement is true.

- If $f(t)$ is a periodic function with period 2π that is sectionally continuous[41] on the interval $(-\pi, \pi)$, then the Fourier series $F(t)$ converges to $f(t)$ at each $t \in (-\pi, \pi)$, where f is continuous. If f is discontinuous at t_0 but possesses left-hand and right-hand derivatives at t_0, then $F(t_0)$ converges to the average value

$$F(t_0) = \frac{f(t_0^-) + f(t_0^+)}{2},$$

  where $f(t_0^-)$ and $f(t_0^+)$ denote the one-sided limits $f(t_0^-) = \lim_{t \to t_0^-} f(t)$ and $f(t_0^+) = \lim_{t \to t_0^+} f(t)$.

For example, the ***square wave function*** defined by

$$f(t) = \begin{cases} -1 & \text{when } -\pi < t < 0, \\ 1 & \text{when } \quad 0 < t < \pi, \end{cases}$$

[41] A function f is sectionally continuous on (a, b) when f has only a finite number of discontinuities in (a, b) and the one-sided limits exist at each point of discontinuity as well as at the end points a and b.

and illustrated in Figure 5.4.2, satisfies these conditions. The value of f at $t = 0$ is irrelevant—it's not even necessary that $f(0)$ be defined.

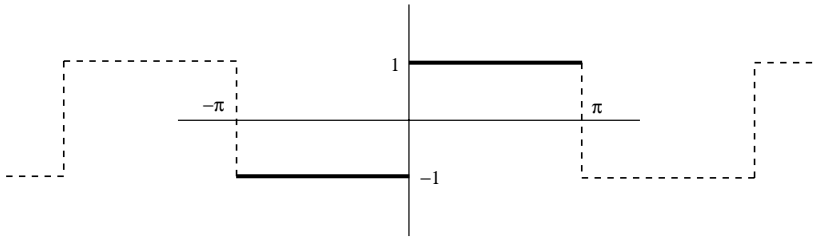


FIGURE 5.4.2

To find the Fourier series expansion for f, compute the coefficients in (5.4.6) as

$$\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos nt\,dt = \frac{1}{\pi}\int_{-\pi}^{0} -\cos nt\,dt + \frac{1}{\pi}\int_{0}^{\pi}\cos nt\,dt \\ &= 0, \\ b_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt\,dt = \frac{1}{\pi}\int_{-\pi}^{0} -\sin nt\,dt + \frac{1}{\pi}\int_{0}^{\pi}\sin nt\,dt \\ &= \frac{2}{n\pi}(1-\cos n\pi) = \begin{cases} 0 & \text{when } n \text{ is even,} \\ 4/n\pi & \text{when } n \text{ is odd,} \end{cases} \end{aligned}$$

so that

$$F(t) = \frac{4}{\pi}\sin t + \frac{4}{3\pi}\sin 3t + \frac{4}{5\pi}\sin 5t + \cdots = \sum_{n=1}^{\infty}\frac{4}{(2n-1)\pi}\sin(2n-1)t.$$

For each $t \in (-\pi, \pi)$, except $t = 0$, it must be the case that $F(t) = f(t)$, and

$$F(0) = \frac{f(0^-) + f(0^+)}{2} = 0.$$

Not only does $F(t)$ agree with $f(t)$ everywhere f is defined, but F also provides a *periodic extension* of f in the sense that the graph of $F(t)$ is the entire square wave depicted in Figure 5.4.2—the values at the points of discontinuity (the jumps) are $F(\pm n\pi) = 0$.

Exercises for section 5.4

5.4.1. Using the standard inner product, determine which of the following pairs are orthogonal vectors in the indicated space.

(a) $\mathbf{x} = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} -2 \\ 2 \\ 2 \end{pmatrix}$ in $\Re^3$,

(b) $\mathbf{x} = \begin{pmatrix} \mathrm{i} \\ 1+\mathrm{i} \\ 2 \\ 1-\mathrm{i} \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 0 \\ 1+\mathrm{i} \\ -2 \\ 1-\mathrm{i} \end{pmatrix}$ in $\mathcal{C}^4$,

(c) $\mathbf{x} = \begin{pmatrix} 1 \\ -2 \\ 3 \\ 4 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 4 \\ 2 \\ -1 \\ 1 \end{pmatrix}$ in $\Re^4$,

(d) $\mathbf{x} = \begin{pmatrix} 1+\mathrm{i} \\ 1 \\ \mathrm{i} \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1-\mathrm{i} \\ -3 \\ -\mathrm{i} \end{pmatrix}$ in $\mathcal{C}^3$,

(e) $\mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ in $\Re^n$.

5.4.2. Find two vectors of unit norm that are orthogonal to $\mathbf{u} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$.

5.4.3. Consider the following set of three vectors.

$$\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix}, \ \mathbf{x}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \ \mathbf{x}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \\ 0 \end{pmatrix} \right\}.$$

(a) Using the standard inner product in $\Re^4$, verify that these vectors are mutually orthogonal.

(b) Find a nonzero vector $\mathbf{x}_4$ such that $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ is a set of mutually orthogonal vectors.

(c) Convert the resulting set into an orthonormal basis for $\Re^4$.

5.4.4. Using the standard inner product, determine the Fourier expansion of $\mathbf{x}$ with respect to $\mathcal{B}$, where

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} \right\}.$$

5.4.5. With respect to the inner product for matrices given by (5.3.2), verify that the set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \right\}$$

is an orthonormal basis for $\Re^{2\times 2}$, and then compute the Fourier expansion of $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ with respect to $\mathcal{B}$.

5.4.6. Determine the angle between $\mathbf{x} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$.

5.4.7. Given an orthonormal basis $\mathcal{B}$ for a space $\mathcal{V}$, explain why the Fourier expansion for $\mathbf{x} \in \mathcal{V}$ is uniquely determined by $\mathcal{B}$.

5.4.8. Explain why the columns of $\mathbf{U}_{n\times n}$ are an orthonormal basis for $\mathcal{C}^n$ if and only if $\mathbf{U}^* = \mathbf{U}^{-1}$. Such matrices are said to be ***unitary***—their properties are studied in a later section.

5.4.9. Matrices with the property $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$ are said to be ***normal.*** Notice that hermitian matrices as well as real symmetric matrices are included in the class of normal matrices. Prove that if $\mathbf{A}$ is normal, then $R(\mathbf{A}) \perp N(\mathbf{A})$—i.e., every vector in $R(\mathbf{A})$ is orthogonal to every vector in $N(\mathbf{A})$. **Hint:** Recall equations (4.5.5) and (4.5.6).

5.4.10. Using the trace inner product described in Example 5.3.1, determine the angle between the following pairs of matrices.

(a) $\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

(b) $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 2 & -2 \\ 2 & 0 \end{pmatrix}$.

5.4.11. Why is the definition for $\cos\theta$ given in (5.4.1) not good for $\mathcal{C}^n$? Explain how to define $\cos\theta$ so that it makes sense in $\mathcal{C}^n$.

5.4.12. If $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ is an orthonormal basis for an inner-product space $\mathcal{V}$, explain why

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_i \langle \mathbf{x}|\mathbf{u}_i\rangle \langle \mathbf{u}_i|\mathbf{y}\rangle$$

holds for every $\mathbf{x}, \mathbf{y} \in \mathcal{V}$.

5.4.13. Consider a real inner-product space, where $\|\star\|^2 = \langle\star|\star\rangle$.

(a) Prove that if $\|\mathbf{x}\| = \|\mathbf{y}\|$, then $(\mathbf{x}+\mathbf{y}) \perp (\mathbf{x}-\mathbf{y})$.

(b) For the standard inner product in $\Re^2$, draw a picture of this. That is, sketch the location of $\mathbf{x}+\mathbf{y}$ and $\mathbf{x}-\mathbf{y}$ for two vectors with equal norms.

5.4.14. Pythagorean Theorem. Let $\mathcal{V}$ be a general inner-product space in which $\|\star\|^2 = \langle\star|\star\rangle$.

(a) When $\mathcal{V}$ is a *real* space, prove that $\mathbf{x} \perp \mathbf{y}$ if and only if $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$. (Something would be wrong if this were not true because this is where the definition of orthogonality originated.)

(b) Construct an example to show that one of the implications in part (a) does not hold when $\mathcal{V}$ is a *complex* space.

(c) When $\mathcal{V}$ is a complex space, prove that $\mathbf{x} \perp \mathbf{y}$ if and only if $\|\alpha\mathbf{x}+\beta\mathbf{y}\|^2 = \|\alpha\mathbf{x}\|^2 + \|\beta\mathbf{y}\|^2$ for all scalars α and β.

5.4.15. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis for an inner-product space $\mathcal{V}$, and let $\mathbf{x} = \sum_i \xi_i \mathbf{u}_i$ be the Fourier expansion of $\mathbf{x} \in \mathcal{V}$.

(a) If $\mathcal{V}$ is a real space, and if θ_i is the angle between $\mathbf{u}_i$ and $\mathbf{x}$, explain why

$$\xi_i = \|\mathbf{x}\| \cos\theta_i.$$

Sketch a picture of this in $\Re^2$ or $\Re^3$ to show why the component $\xi_i\mathbf{u}_i$ represents the orthogonal projection of $\mathbf{x}$ onto the line determined by $\mathbf{u}_i$, and thus illustrate the fact that a Fourier expansion is nothing more than simply resolving $\mathbf{x}$ into mutually orthogonal components.

(b) Derive ***Parseval's identity***,[42] which says $\sum_{i=1}^n |\xi_i|^2 = \|\mathbf{x}\|^2$.

5.4.16. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ be an orthonormal set in an n-dimensional inner-product space $\mathcal{V}$. Derive ***Bessel's inequality***,[43] which says that if $\mathbf{x} \in \mathcal{V}$ and $\xi_i = \langle\mathbf{u}_i|\mathbf{x}\rangle$, then

$$\sum_{i=1}^k |\xi_i|^2 \le \|\mathbf{x}\|^2.$$

Explain why equality holds if and only if $\mathbf{x} \in span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$. **Hint:** Consider $\|\mathbf{x} - \sum_{i=1}^k \xi_i\mathbf{u}_i\|^2$.

---

[42] This result appeared in the second of the five mathematical publications by Marc-Antoine Parseval des Chênes (1755–1836). Parseval was a royalist who had to flee from France when Napoleon ordered his arrest for publishing poetry against the regime.

[43] This inequality is named in honor of the German astronomer and mathematician Friedrich Wilhelm Bessel (1784–1846), who devoted his life to understanding the motions of the stars. In the process he introduced several useful mathematical ideas.

5.4.17. Construct an example using the standard inner product in $\Re^n$ to show that two vectors $\mathbf{x}$ and $\mathbf{y}$ can have an angle between them that is close to $\pi/2$ without $\mathbf{x}^T\mathbf{y}$ being close to 0. **Hint:** Consider n to be large, and use the vector $\mathbf{e}$ of all 1's for one of the vectors.

5.4.18. It was demonstrated in Example 5.4.3 that $\mathbf{y}$ is linearly correlated with $\mathbf{x}$ in the sense that $\mathbf{y} \approx \beta_0\mathbf{e} + \beta_1\mathbf{x}$ if and only if the standardization vectors $\mathbf{z_x}$ and $\mathbf{z_y}$ are "close" in the sense that they are almost on the same line in $\Re^n$. Explain why simply measuring $\|\mathbf{z_x} - \mathbf{z_y}\|_2$ does not always gauge the degree of linear correlation.

5.4.19. Let θ be the angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ from a real inner-product space.

(a) Prove that $\cos\theta = 1$ if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha > 0$.

(b) Prove that $\cos\theta = -1$ if and only if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha < 0$.

Hint: Use the generalization of Exercise 5.1.9.

5.4.20. With respect to the orthonormal set

$$\mathcal{B} = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{\cos t}{\sqrt{\pi}}, \frac{\cos 2t}{\sqrt{\pi}}, \ldots, \frac{\sin t}{\sqrt{\pi}}, \frac{\sin 2t}{\sqrt{\pi}}, \frac{\sin 3t}{\sqrt{\pi}}, \ldots \right\},$$

determine the Fourier series expansion of the ***saw-toothed function*** defined by $f(t) = t$ for $-\pi < t < \pi$. The periodic extension of this function is depicted in Figure 5.4.3.

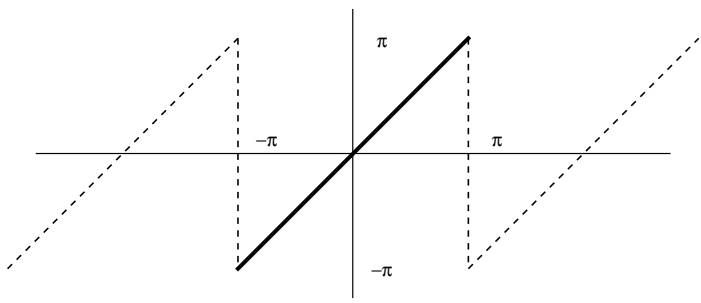


FIGURE 5.4.3

5.5 GRAM–SCHMIDT PROCEDURE

As discussed in §5.4, orthonormal bases possess significant advantages over bases that are not orthonormal. The spaces $\Re^n$ and $\mathcal{C}^n$ clearly possess orthonormal bases (e.g., the standard basis), but what about other spaces? Does every finite-dimensional space possess an orthonormal basis, and, if so, how can one be produced? The ***Gram–Schmidt***[44] orthogonalization procedure developed below answers these questions.

Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be an arbitrary basis (not necessarily orthonormal) for an n-dimensional inner-product space $\mathcal{S}$, and remember that $\|\star\| = \langle \star | \star \rangle^{1/2}$.

Objective: Use $\mathcal{B}$ to construct an orthonormal basis $\mathcal{O} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ for $\mathcal{S}$.

Strategy: Construct $\mathcal{O}$ sequentially so that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is an orthonormal basis for $\mathcal{S}_k = span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for $k = 1, \ldots, n$.

For $k = 1$, simply take $\mathbf{u}_1 = \mathbf{x}_1 / \|\mathbf{x}_1\|$. It's clear that $\mathcal{O}_1 = \{\mathbf{u}_1\}$ is an orthonormal set whose span agrees with that of $\mathcal{S}_1 = \{\mathbf{x}_1\}$. Now reason inductively. Suppose that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is an orthonormal basis for $\mathcal{S}_k = span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$, and consider the problem of finding one additional vector $\mathbf{u}_{k+1}$ such that $\mathcal{O}_{k+1} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \mathbf{u}_{k+1}\}$ is an orthonormal basis for $\mathcal{S}_{k+1} = span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \mathbf{x}_{k+1}\}$. For this to hold, the Fourier expansion (p. 299) of $\mathbf{x}_{k+1}$ with respect to $\mathcal{O}_{k+1}$ must be

$$\mathbf{x}_{k+1} = \sum_{i=1}^{k+1} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \, \mathbf{u}_i,$$

which in turn implies that

$$\mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \, \mathbf{u}_i}{\langle \mathbf{u}_{k+1} | \mathbf{x}_{k+1} \rangle}. \tag{5.5.1}$$

Since $\|\mathbf{u}_{k+1}\| = 1$, it follows from (5.5.1) that

$$|\,\langle \mathbf{u}_{k+1} | \mathbf{x}_{k+1} \rangle\,| = \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \, \mathbf{u}_i \right\|,$$

[44] Jorgen P. Gram (1850–1916) was a Danish actuary who implicitly presented the essence of orthogonalization procedure in 1883. Gram was apparently unaware that Pierre-Simon Laplace (1749–1827) had earlier used the method. Today, Gram is remembered primarily for his development of this process, but in earlier times his name was also associated with the matrix product $\mathbf{A}^*\mathbf{A}$ that historically was referred to as the *Gram matrix* of $\mathbf{A}$.

Erhard Schmidt (1876–1959) was a student of Hermann Schwarz (of CBS inequality fame) and the great German mathematician David Hilbert. Schmidt explicitly employed the orthogonalization process in 1907 in his study of integral equations, which in turn led to the development of what are now called *Hilbert spaces*. Schmidt made significant use of the orthogonalization process to develop the geometry of Hilbert Spaces, and thus it came to bear Schmidt's name.

so $\langle \mathbf{u}_{k+1} | \mathbf{x}_{k+1} \rangle = \mathrm{e}^{\mathrm{i}\theta} \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|$ for some $0 \le \theta < 2\pi$, and

$$\mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i}{\mathrm{e}^{\mathrm{i}\theta} \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|}.$$

Since the value of θ in the scalar $\mathrm{e}^{\mathrm{i}\theta}$ neither affects $span\,\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k+1}\}$ nor the facts that $\|\mathbf{u}_{k+1}\| = 1$ and $\langle \mathbf{u}_{k+1} | \mathbf{u}_i \rangle = 0$ for all $i \le k$, we can arbitrarily define $\mathbf{u}_{k+1}$ to be the vector corresponding to the $\theta = 0$ or, equivalently, $\mathrm{e}^{\mathrm{i}\theta} = 1$. For the sake of convenience, let

$$\nu_{k+1} = \left\| \mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\|$$

so that we can write

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_{k+1} = \frac{\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i}{\nu_{k+1}} \text{ for } k > 0. \tag{5.5.2}$$

This sequence of vectors is called the ***Gram–Schmidt sequence.*** A straightforward induction argument proves that $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is indeed an orthonormal basis for $span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for each $k = 1, 2, \ldots$. Details are called for in Exercise 5.5.7.

The orthogonalization procedure defined by (5.5.2) is valid for any inner-product space, but if we concentrate on subspaces of $\Re^m$ or $\mathcal{C}^m$ with the standard inner product and euclidean norm, then we can formulate (5.5.2) in terms of matrices. Suppose that $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis for an n-dimensional subspace $\mathcal{S}$ of $\mathcal{C}^{m\times 1}$ so that the Gram–Schmidt sequence (5.5.2) becomes

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_k = \frac{\mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k)\, \mathbf{u}_i}{\left\| \mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k)\, \mathbf{u}_i \right\|} \quad \text{for } k = 2, 3, \ldots, n. \tag{5.5.3}$$

To express this in matrix notation, set

$$\mathbf{U}_1 = \mathbf{0}_{m\times 1} \quad \text{and} \quad \mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big)_{m\times k-1} \quad \text{for } k > 1,$$

and notice that

$$\mathbf{U}_k^* \mathbf{x}_k = \begin{pmatrix} \mathbf{u}_1^* \mathbf{x}_k \\ \mathbf{u}_2^* \mathbf{x}_k \\ \vdots \\ \mathbf{u}_{k-1}^* \mathbf{x}_k \end{pmatrix} \quad \text{and} \quad \mathbf{U}_k \mathbf{U}_k^* \mathbf{x}_k = \sum_{i=1}^{k-1} \mathbf{u}_i\, (\mathbf{u}_i^* \mathbf{x}_k) = \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k)\, \mathbf{u}_i.$$

Since

$$\mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k)\, \mathbf{u}_i = \mathbf{x}_k - \mathbf{U}_k \mathbf{U}_k^* \mathbf{x}_k = (\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*)\, \mathbf{x}_k,$$

the vectors in (5.5.3) can be concisely written as

$$\mathbf{u}_k = \frac{(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*)\, \mathbf{x}_k}{\|(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*)\, \mathbf{x}_k\|} \quad \text{for } k = 1, 2, \ldots, n.$$

Below is a summary.

Gram–Schmidt Orthogonalization Procedure

If $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis for a general inner-product space $\mathcal{S}$, then the ***Gram–Schmidt sequence*** defined by

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \quad \text{and} \quad \mathbf{u}_k = \frac{\mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \, \mathbf{u}_i}{\left\| \mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \, \mathbf{u}_i \right\|} \quad \text{for } k = 2, \ldots, n$$

is an orthonormal basis for $\mathcal{S}$. When $\mathcal{S}$ is an n-dimensional subspace of $\mathcal{C}^{m\times 1}$, the Gram–Schmidt sequence can be expressed as

$$\mathbf{u}_k = \frac{(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*)\, \mathbf{x}_k}{\|(\mathbf{I} - \mathbf{U}_k \mathbf{U}_k^*)\, \mathbf{x}_k\|} \quad \text{for} \quad k = 1, 2, \ldots, n \tag{5.5.4}$$

in which $\mathbf{U}_1 = \mathbf{0}_{m\times 1}$ and $\mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big)_{m\times k-1}$ for $k > 1$.

Example 5.5.1

Classical Gram–Schmidt Algorithm. The following formal algorithm is the straightforward or "classical" implementation of the Gram–Schmidt procedure. Interpret $\mathbf{a} \leftarrow \mathbf{b}$ to mean that "$\mathbf{a}$ is defined to be (or overwritten by) $\mathbf{b}$."

For $k = 1$:

$$\mathbf{u}_1 \leftarrow \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$$

For $k > 1$:

$$\mathbf{u}_k \leftarrow \mathbf{x}_k - \sum_{i=1}^{k-1} (\mathbf{u}_i^* \mathbf{x}_k) \mathbf{u}_i$$

$$\mathbf{u}_k \leftarrow \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}$$

(See Exercise 5.5.10 for other formulations of the Gram–Schmidt algorithm.)

Problem: Use the classical formulation of the Gram–Schmidt procedure given above to find an orthonormal basis for the space spanned by the following three linearly independent vectors.

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 3 \\ 1 \\ 1 \\ -1 \end{pmatrix}.$$

Solution:

$$k=1: \quad \mathbf{u}_1 \leftarrow \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\\-1\end{pmatrix}$$

$$k=2: \quad \mathbf{u}_2 \leftarrow \mathbf{x}_2 - (\mathbf{u}_1^T\mathbf{x}_2)\mathbf{u}_1 = \begin{pmatrix}0\\2\\0\\0\end{pmatrix}, \quad \mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}$$

$$k=3: \quad \mathbf{u}_3 \leftarrow \mathbf{x}_3 - (\mathbf{u}_1^T\mathbf{x}_3)\mathbf{u}_1 - (\mathbf{u}_2^T\mathbf{x}_3)\mathbf{u}_2 = \begin{pmatrix}1\\0\\1\\1\end{pmatrix}, \quad \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\0\\1\\1\end{pmatrix}$$

Thus

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\\-1\end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\0\\1\\1\end{pmatrix}$$

is the desired orthonormal basis.

The Gram–Schmidt process frequently appears in the disguised form of a matrix factorization. To see this, let $\mathbf{A}_{m\times n} = \big(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\big)$ be a matrix with linearly independent columns. When Gram–Schmidt is applied to the columns of $\mathbf{A}$, the result is an orthonormal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ for $R(\mathbf{A})$, where

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\nu_1} \quad \text{and} \quad \mathbf{q}_k = \frac{\mathbf{a}_k - \sum_{i=1}^{k-1}\langle \mathbf{q}_i | \mathbf{a}_k\rangle\, \mathbf{q}_i}{\nu_k} \quad \text{for } k = 2, 3, \ldots, n,$$

where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \left\|\mathbf{a}_k - \sum_{i=1}^{k-1}\langle \mathbf{q}_i | \mathbf{a}_k\rangle\, \mathbf{q}_i\right\|$ for $k > 1$. The above relationships can be rewritten as

$$\mathbf{a}_1 = \nu_1\mathbf{q}_1 \quad \text{and} \quad \mathbf{a}_k = \langle \mathbf{q}_1 | \mathbf{a}_k\rangle\, \mathbf{q}_1 + \cdots + \langle \mathbf{q}_{k-1} | \mathbf{a}_k\rangle\, \mathbf{q}_{k-1} + \nu_k\mathbf{q}_k \quad \text{for } k > 1,$$

which in turn can be expressed in matrix form by writing

$$\big(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\big) = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\big)\begin{pmatrix} \nu_1 & \langle \mathbf{q}_1|\mathbf{a}_2\rangle & \langle \mathbf{q}_1|\mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_1|\mathbf{a}_n\rangle \\ 0 & \nu_2 & \langle \mathbf{q}_2|\mathbf{a}_3\rangle & \cdots & \langle \mathbf{q}_2|\mathbf{a}_n\rangle \\ 0 & 0 & \nu_3 & \cdots & \langle \mathbf{q}_3|\mathbf{a}_n\rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \nu_n \end{pmatrix}.$$

This says that it's possible to factor a matrix with independent columns as $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$, where the columns of $\mathbf{Q}$ are an orthonormal basis for $R(\mathbf{A})$ and $\mathbf{R}$ is an upper-triangular matrix with positive diagonal elements.

The factorization $\mathbf{A} = \mathbf{QR}$ is called the ***QR factorization*** for $\mathbf{A}$, and it is uniquely determined by $\mathbf{A}$ (Exercise 5.5.8). When $\mathbf{A}$ and $\mathbf{Q}$ are not square, some authors emphasize the point by calling $\mathbf{A} = \mathbf{QR}$ the *rectangular* QR factorization—the case when $\mathbf{A}$ and $\mathbf{Q}$ are square is further discussed on p. 345. Below is a summary of the above observations.

QR Factorization

Every matrix $\mathbf{A}_{m\times n}$ with linearly independent columns can be uniquely factored as $\mathbf{A} = \mathbf{QR}$ in which the columns of $\mathbf{Q}_{m\times n}$ are an orthonormal basis for $R(\mathbf{A})$ and $\mathbf{R}_{n\times n}$ is an upper-triangular matrix with positive diagonal entries.

- The QR factorization is the complete "road map" of the Gram–Schmidt process because the columns of $\mathbf{Q} = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\big)$ are the result of applying the Gram–Schmidt procedure to the columns of $\mathbf{A} = \big(\mathbf{a}_1 \,|\, \mathbf{a}_2 \,|\, \cdots \,|\, \mathbf{a}_n\big)$ and $\mathbf{R}$ is given by

$$\mathbf{R} = \begin{pmatrix} \nu_1 & \mathbf{q}_1^*\mathbf{a}_2 & \mathbf{q}_1^*\mathbf{a}_3 & \cdots & \mathbf{q}_1^*\mathbf{a}_n \\ 0 & \nu_2 & \mathbf{q}_2^*\mathbf{a}_3 & \cdots & \mathbf{q}_2^*\mathbf{a}_n \\ 0 & 0 & \nu_3 & \cdots & \mathbf{q}_3^*\mathbf{a}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \nu_n \end{pmatrix},$$

where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \left\|\mathbf{a}_k - \sum_{i=1}^{k-1} \langle \mathbf{q}_i | \mathbf{a}_k \rangle \, \mathbf{q}_i\right\|$ for $k > 1$.

Example 5.5.2

Problem: Determine the QR factors of

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}.$$

Solution: Using the standard inner product for $\Re^n$, apply the Gram–Schmidt procedure to the columns of $\mathbf{A}$ by setting

$$\mathbf{q}_1 = \frac{\mathbf{a}_1}{\nu_1} \quad \text{and} \quad \mathbf{q}_k = \frac{\mathbf{a}_k - \sum_{i=1}^{k-1} \left(\mathbf{q}_i^T \mathbf{a}_k\right) \mathbf{q}_i}{\nu_k} \quad \text{for } k = 2, 3,$$

where $\nu_1 = \|\mathbf{a}_1\|$ and $\nu_k = \left\|\mathbf{a}_k - \sum_{i=1}^{k-1} \left(\mathbf{q}_i^T \mathbf{a}_k\right) \mathbf{q}_i\right\|$. The computation of these quantities can be organized as follows.

$$k=1: \quad r_{11} \leftarrow \|\mathbf{a}_1\| = 5 \qquad \text{and} \qquad \mathbf{q}_1 \leftarrow \frac{\mathbf{a}_1}{r_{11}} = \begin{pmatrix} 0 \\ 3/5 \\ 4/5 \end{pmatrix}$$

$$k=2: \quad r_{12} \leftarrow \mathbf{q}_1^T\mathbf{a}_2 = 25$$

$$\mathbf{q}_2 \leftarrow \mathbf{a}_2 - r_{12}\mathbf{q}_1 = \begin{pmatrix} -20 \\ 12 \\ -9 \end{pmatrix}$$

$$r_{22} \leftarrow \|\mathbf{q}_2\| = 25 \text{ and } \mathbf{q}_2 \leftarrow \frac{\mathbf{q}_2}{r_{22}} = \frac{1}{25}\begin{pmatrix} -20 \\ 12 \\ -9 \end{pmatrix}$$

$$k=3: \quad r_{13} \leftarrow \mathbf{q}_1^T\mathbf{a}_3 = -4 \text{ and } r_{23} \leftarrow \mathbf{q}_2^T\mathbf{a}_3 = 10$$

$$\mathbf{q}_3 \leftarrow \mathbf{a}_3 - r_{13}\mathbf{q}_1 - r_{23}\mathbf{q}_2 = \frac{2}{5}\begin{pmatrix} -15 \\ -16 \\ 12 \end{pmatrix}$$

$$r_{33} \leftarrow \|\mathbf{q}_3\| = 10 \text{ and } \mathbf{q}_3 \leftarrow \frac{\mathbf{q}_3}{r_{33}} = \frac{1}{25}\begin{pmatrix} -15 \\ -16 \\ 12 \end{pmatrix}$$

Therefore,

$$\mathbf{Q} = \frac{1}{25}\begin{pmatrix} 0 & -20 & -15 \\ 15 & 12 & -16 \\ 20 & -9 & 12 \end{pmatrix} \qquad \text{and} \qquad \mathbf{R} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

We now have two important matrix factorizations, namely, the LU factorization, discussed in §3.10 on p. 141 and the QR factorization. They are not the same, but some striking analogies exist.

- Each factorization represents a reduction to upper-triangular form—LU by Gaussian elimination, and QR by Gram–Schmidt. In particular, the LU factorization is the complete "road map" of Gaussian elimination applied to a square nonsingular matrix, whereas QR is the complete road map of Gram–Schmidt applied to a matrix with linearly independent columns.
- When they exist, both factorizations $\mathbf{A} = \mathbf{LU}$ and $\mathbf{A} = \mathbf{QR}$ are uniquely determined by $\mathbf{A}$.
- Once the LU factors (assuming they exist) of a nonsingular matrix $\mathbf{A}$ are known, the solution of $\mathbf{Ax} = \mathbf{b}$ is easily computed—solve $\mathbf{Ly} = \mathbf{b}$ by forward substitution, and then solve $\mathbf{Ux} = \mathbf{y}$ by back substitution (see p. 146). The QR factors can be used in a similar manner. If $\mathbf{A} \in \Re^{n\times n}$ is nonsingular, then $\mathbf{Q}^T = \mathbf{Q}^{-1}$ (because $\mathbf{Q}$ has orthonormal columns), so $\mathbf{Ax} = \mathbf{b} \iff \mathbf{QRx} = \mathbf{b} \iff \mathbf{Rx} = \mathbf{Q}^T\mathbf{b}$, which is also a triangular system that is solved by back substitution.

While the LU and QR factors can be used in more or less the same way to solve nonsingular systems, things are different for singular and rectangular cases because $\mathbf{Ax} = \mathbf{b}$ might be inconsistent, in which case a least squares solution as described in §4.6, (p. 223) may be desired. Unfortunately, the LU factors of $\mathbf{A}$ don't exist when $\mathbf{A}$ is rectangular. And even if $\mathbf{A}$ is square and has an LU factorization, the LU factors of $\mathbf{A}$ are not much help in solving the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ that produces least squares solutions. But the QR factors of $\mathbf{A}_{m\times n}$ always exist as long as $\mathbf{A}$ has linearly independent columns, and, as demonstrated in the following example, the QR factors provide the least squares solution of an inconsistent system in exactly the same way as they provide the solution of a consistent system.

Example 5.5.3

Application to the Least Squares Problem. If $\mathbf{Ax} = \mathbf{b}$ is a possibly inconsistent (real) system, then, as discussed on p. 226, the set of all least squares solutions is the set of solutions to the system of normal equations

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}. \tag{5.5.5}$$

But computing $\mathbf{A}^T\mathbf{A}$ and then performing an LU factorization of $\mathbf{A}^T\mathbf{A}$ to solve (5.5.5) is generally not advisable. First, it's inefficient and, second, as pointed out in Example 4.5.1, computing $\mathbf{A}^T\mathbf{A}$ with floating-point arithmetic can result in a loss of significant information. The QR approach doesn't suffer from either of these objections. Suppose that $rank\,(\mathbf{A}_{m\times n}) = n$ (so that there is a unique least squares solution), and let $\mathbf{A} = \mathbf{QR}$ be the QR factorization. Because the columns of $\mathbf{Q}$ are an orthonormal set, it follows that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$, so

$$\mathbf{A}^T\mathbf{A} = (\mathbf{QR})^T(\mathbf{QR}) = \mathbf{R}^T\mathbf{Q}^T\mathbf{QR} = \mathbf{R}^T\mathbf{R}. \tag{5.5.6}$$

Consequently, the normal equations (5.5.5) can be written as

$$\mathbf{R}^T\mathbf{Rx} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}. \tag{5.5.7}$$

But $\mathbf{R}^T$ is nonsingular (it is triangular with positive diagonal entries), so (5.5.7) simplifies to become

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{5.5.8}$$

This is just an upper-triangular system that is efficiently solved by back substitution. In other words, most of the work involved in solving the least squares problem is in computing the QR factorization of $\mathbf{A}$. Finally, notice that

$$\mathbf{x} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{b} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}$$

is the solution of $\mathbf{Ax} = \mathbf{b}$ when the system is consistent as well as the least squares solution when the system is inconsistent (see p. 214). That is, with the QR approach, it makes no difference whether or not $\mathbf{Ax} = \mathbf{b}$ is consistent because in both cases things boil down to solving the same equation—namely, (5.5.8). Below is a formal summary.

Linear Systems and the QR Factorization

If $rank\,(\mathbf{A}_{m\times n}) = n$, and if $\mathbf{A} = \mathbf{QR}$ is the QR factorization, then the solution of the nonsingular triangular system

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b} \tag{5.5.9}$$

is either the solution or the least squares solution of $\mathbf{Ax} = \mathbf{b}$ depending on whether or not $\mathbf{Ax} = \mathbf{b}$ is consistent.

It's worthwhile to reemphasize that the QR approach to the least squares problem obviates the need to explicitly compute the product $\mathbf{A}^T\mathbf{A}$. But if $\mathbf{A}^T\mathbf{A}$ is ever needed, it is retrievable from the factorization $\mathbf{A}^T\mathbf{A} = \mathbf{R}^T\mathbf{R}$. In fact, this is the ***Cholesky factorization*** of $\mathbf{A}^T\mathbf{A}$ as discussed in Example 3.10.7, p. 154.

The Gram–Schmidt procedure is a powerful theoretical tool, but it's not a good numerical algorithm when implemented in the straightforward or "classical" sense. When floating-point arithmetic is used, the classical Gram–Schmidt algorithm applied to a set of vectors that is not already close to being an orthogonal set can produce a set of vectors that is far from being an orthogonal set. To see this, consider the following example.

Example 5.5.4

Problem: Using 3-digit floating-point arithmetic, apply the classical Gram–Schmidt algorithm to the set

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}.$$

Solution:

$k = 1$: $\quad fl\,\|\mathbf{x}_1\| = 1$, so $\mathbf{u}_1 \leftarrow \mathbf{x}_1$.

$k = 2$: $\quad fl\,(\mathbf{u}_1^T\mathbf{x}_2) = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{x}_2 - (\mathbf{u}_1^T\mathbf{x}_2)\,\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 \leftarrow fl\left(\frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}\right) = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

$k = 3$: $\quad fl\,(\mathbf{u}_1^T\mathbf{x}_3) = 1$ and $fl\,(\mathbf{u}_2^T\mathbf{x}_3) = -10^{-3}$, so

$$\mathbf{u}_3 \leftarrow \mathbf{x}_3 - (\mathbf{u}_1^T\mathbf{x}_3)\mathbf{u}_1 - (\mathbf{u}_2^T\mathbf{x}_3)\mathbf{u}_2 = \begin{pmatrix} 0 \\ -10^{-3} \\ -10^{-3} \end{pmatrix} \text{ and } \mathbf{u}_3 \leftarrow fl\left(\frac{\mathbf{u}_3}{\|\mathbf{u}_3\|}\right) = \begin{pmatrix} 0 \\ -.709 \\ -.709 \end{pmatrix}.$$

Therefore, classical Gram–Schmidt with 3-digit arithmetic returns

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ -.709 \\ -.709 \end{pmatrix}, \tag{5.5.10}$$

which is unsatisfactory because $\mathbf{u_2}$ and $\mathbf{u}_3$ are far from being orthogonal.

It's possible to improve the numerical stability of the orthogonalization process by rearranging the order of the calculations. Recall from (5.5.4) that

$$\mathbf{u}_k = \frac{(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k}{\|(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*)\,\mathbf{x}_k\|}, \quad \text{where} \quad \mathbf{U}_1 = \mathbf{0} \text{ and } \mathbf{U}_k = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\big).$$

If $\mathbf{E}_1 = \mathbf{I}$ and $\mathbf{E}_i = \mathbf{I} - \mathbf{u}_{i-1}\mathbf{u}_{i-1}^*$ for $i > 1$, then the orthogonality of the $\mathbf{u}_i$'s insures that

$$\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1 = \mathbf{I} - \mathbf{u}_1\mathbf{u}_1^* - \mathbf{u}_2\mathbf{u}_2^* - \cdots - \mathbf{u}_{k-1}\mathbf{u}_{k-1}^* = \mathbf{I} - \mathbf{U}_k\mathbf{U}_k^*,$$

so the Gram–Schmidt sequence can also be expressed as

$$\mathbf{u}_k = \frac{\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k}{\|\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k\|} \quad \text{for } k = 1, 2, \ldots, n.$$

This means that the Gram–Schmidt sequence can be generated as follows:

$$\begin{aligned}
\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} &\xrightarrow{\text{Normalize 1-st}} \{\mathbf{u}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \\
&\xrightarrow{\text{Apply } \mathbf{E}_2} \{\mathbf{u}_1,\ \mathbf{E}_2\mathbf{x}_2,\ \mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Normalize 2-nd}} \{\mathbf{u}_1, \mathbf{u}_2,\ \mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Apply } \mathbf{E}_3} \{\mathbf{u}_1, \mathbf{u}_2,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_3,\ \ldots,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_n\} \\
&\xrightarrow{\text{Normalize 3-rd}} \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_4, \ldots,\ \mathbf{E}_3\mathbf{E}_2\mathbf{x}_n\}, \\
&\qquad \text{etc.}
\end{aligned}$$

While there is no theoretical difference, this "modified" algorithm is numerically more stable than the classical algorithm when floating-point arithmetic is used. The kth step of the classical algorithm alters only the kth vector, but the kth step of the modified algorithm "updates" all vectors from the kth through the last, and conditioning the unorthogonalized tail in this way makes a difference.

Modified Gram–Schmidt Algorithm

For a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathcal{C}^{m\times 1}$, the Gram–Schmidt sequence given on p. 309 can be alternately described as

$$\mathbf{u}_k = \frac{\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k}{\|\mathbf{E}_k \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{x}_k\|} \text{ with } \mathbf{E}_1 = \mathbf{I}, \;\; \mathbf{E}_i = \mathbf{I} - \mathbf{u}_{i-1}\mathbf{u}_{i-1}^* \text{ for } i > 1,$$

and this sequence is generated by the following algorithm.

For $k = 1$: $\quad \mathbf{u}_1 \leftarrow \mathbf{x}_1 / \|\mathbf{x}_1\| \quad$ and $\quad \mathbf{u}_j \leftarrow \mathbf{x}_j$ for $j = 2, 3, \ldots, n$

For $k > 1$: $\quad \mathbf{u}_j \leftarrow \mathbf{E}_k\mathbf{u}_j = \mathbf{u}_j - \left(\mathbf{u}_{k-1}^*\mathbf{u}_j\right)\mathbf{u}_{k-1}$ for $j = k, k+1, \ldots, n$

$\qquad\qquad\quad \mathbf{u}_k \leftarrow \mathbf{u}_k / \|\mathbf{u}_k\|$

(An alternate implementation is given in Exercise 5.5.10.)

To see that the modified version of Gram–Schmidt can indeed make a difference when floating-point arithmetic is used, consider the following example.

Example 5.5.5

Problem: Use 3-digit floating-point arithmetic, and apply the modified Gram–Schmidt algorithm to the set given in Example 5.5.4 (p. 314), and then compare the results of the modified algorithm with those of the classical algorithm.

Solution: $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}.$

$k = 1$: $\quad fl\,\|\mathbf{x}_1\| = 1$, so $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \leftarrow \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$.

$k = 2$: $\quad fl\left(\mathbf{u}_1^T\mathbf{u}_2\right) = 1$ and $fl\left(\mathbf{u}_1^T\mathbf{u}_3\right) = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{u}_2 - \left(\mathbf{u}_1^T\mathbf{u}_2\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix}, \quad \mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_1^T\mathbf{u}_3\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ -10^{-3} \\ 0 \end{pmatrix},$$

and

$$\mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

$k = 3$: $\quad \mathbf{u}_2^T\mathbf{u}_3 = 0$, so

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_2^T\mathbf{u}_3\right)\mathbf{u}_2 = \begin{pmatrix} 0 \\ -10^{-3} \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}.$$

Thus the modified Gram–Schmidt algorithm produces

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 10^{-3} \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \tag{5.5.11}$$

which is as good as one can expect using 3-digit arithmetic. Comparing (5.5.11) with the result (5.5.10) obtained in Example 5.5.4 illuminates the advantage possessed by modified Gram–Schmidt algorithm over the classical algorithm.

Below is a summary of some facts concerning the modified Gram–Schmidt algorithm compared with the classical implementation.

Summary

- When the Gram–Schmidt procedures (classical or modified) are applied to the columns of $\mathbf{A}$ using exact arithmetic, each produces an orthonormal basis for $R(\mathbf{A})$.
- For computing a QR factorization in floating-point arithmetic, the modified algorithm produces results that are at least as good as and often better than the classical algorithm, but the modified algorithm is not unconditionally stable—there are situations in which it fails to produce a set of columns that are nearly orthogonal.
- For solving the least square problem with floating-point arithmetic, the modified procedure is a numerically stable algorithm in the sense that the method described in Example 5.5.3 returns a result that is the exact solution of a nearby least squares problem. However, the Householder method described on p. 346 is just as stable and needs slightly fewer arithmetic operations.

Exercises for section 5.5

5.5.1. Let $\mathcal{S} = span\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} 2 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \begin{pmatrix} -1 \\ 2 \\ 2 \\ 1 \end{pmatrix} \right\}$.

(a) Use the classical Gram–Schmidt algorithm (with exact arithmetic) to determine an orthonormal basis for $\mathcal{S}$.

(b) Verify directly that the Gram–Schmidt sequence produced in part (a) is indeed an orthonormal basis for $\mathcal{S}$.

(c) Repeat part (a) using the modified Gram–Schmidt algorithm, and compare the results.

5.5.2. Use the Gram–Schmidt procedure to find an orthonormal basis for the four fundamental subspaces of $\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 & -1 \\ 2 & -4 & 6 & -2 \\ 3 & -6 & 9 & -3 \end{pmatrix}$.

5.5.3. Apply the Gram–Schmidt procedure with the standard inner product for $\mathcal{C}^3$ to $\left\{ \begin{pmatrix} \mathrm{i} \\ \mathrm{i} \\ \mathrm{i} \end{pmatrix}, \begin{pmatrix} 0 \\ \mathrm{i} \\ \mathrm{i} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \mathrm{i} \end{pmatrix} \right\}$.

5.5.4. Explain what happens when the Gram–Schmidt process is applied to an orthonormal set of vectors.

5.5.5. Explain what happens when the Gram–Schmidt process is applied to a linearly dependent set of vectors.

5.5.6. Let $\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & -3 \\ 0 & 1 & 1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

(a) Determine the rectangular QR factorization of $\mathbf{A}$.

(b) Use the QR factors from part (a) to determine the least squares solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$.

5.5.7. Given a linearly independent set of vectors $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in an inner-product space, let $\mathcal{S}_k = span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for $k = 1, 2, \ldots, n$. Give an induction argument to prove that if $\mathcal{O}_k = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is the Gram–Schmidt sequence defined in (5.5.2), then $\mathcal{O}_k$ is indeed an orthonormal basis for $\mathcal{S}_k = span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ for each $k = 1, 2, \ldots, n$.

5.5.8. Prove that if $rank\,(\mathbf{A}_{m\times n}) = n$, then the rectangular QR factorization of $\mathbf{A}$ is unique. That is, if $\mathbf{A} = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q}_{m\times n}$ has orthonormal columns and $\mathbf{R}_{n\times n}$ is upper triangular with positive diagonal entries, then $\mathbf{Q}$ and $\mathbf{R}$ are unique. **Hint:** Recall Example 3.10.7, p. 154.

5.5.9. (a) Apply classical Gram–Schmidt with 3-digit floating-point arithmetic to $\left\{ \mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 1 \\ 10^{-3} \\ 0 \end{pmatrix} \right\}$. You may assume that $fl\left(\sqrt{2}\right) = 1.41$.

(b) Again using 3-digit floating-point arithmetic, apply the modified Gram–Schmidt algorithm to $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, and compare the result with that of part (a).

5.5.10. Depending on how the inner products r_{ij} are defined, verify that the following code implements both the classical and modified Gram–Schmidt algorithms applied to a set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$.

```
For j = 1 to n
  u_j ⟵ x_j
        For i = 1 to j − 1
          r_ij ⟵ { ⟨u_i|x_j⟩  (classical Gram–Schmidt)
                 { ⟨u_i|u_j⟩  (modified Gram–Schmidt)
          u_j ⟵ u_j − r_ij u_i
        End
  r_jj ⟵ ‖u_j‖
        If r_jj = 0
          quit  (because x_j ∈ span{x_1, x_2, ..., x_{j−1}})
        Else u_j ⟵ u_j / r_jj
End
```

If exact arithmetic is used, will the inner products r_{ij} be the same for both implementations?

5.5.11. Let $\mathcal{V}$ be the inner-product space of real-valued continuous functions defined on the interval $[-1, 1]$, where the inner product is defined by

$$\langle f|g\rangle = \int_{-1}^{1} f(x)g(x)dx,$$

and let $\mathcal{S}$ be the subspace of $\mathcal{V}$ that is spanned by the three linearly independent polynomials $q_0 = 1, \quad q_1 = x, \quad q_2 = x^2$.

(a) Use the Gram–Schmidt process to determine an orthonormal set of polynomials $\{p_0, p_1, p_2\}$ that spans $\mathcal{S}$. These polynomials are the first three normalized ***Legendre***[45] ***polynomials.***

(b) Verify that p_n satisfies ***Legendre's differential equation***

$$(1 - x^2)y'' - 2xy' + n(n+1)y = 0$$

for $n = 0, 1, 2$. This equation and its solutions are of considerable importance in applied mathematics.

[45] Adrien–Marie Legendre (1752–1833) was one of the most eminent French mathematicians of the eighteenth century. His primary work in higher mathematics concerned number theory and the study of elliptic functions. But he was also instrumental in the development of the theory of least squares, and some people believe that Legendre should receive the credit that is often afforded to Gauss for the introduction of the method of least squares. Like Gauss and many other successful mathematicians, Legendre spent substantial time engaged in diligent and painstaking computation. It is reported that in 1824 Legendre refused to vote for the government's candidate for Institut National, so his pension was stopped, and he died in poverty.

5.6 UNITARY AND ORTHOGONAL MATRICES

The purpose of this section is to examine square matrices whose columns (or rows) are orthonormal. The standard inner product and the euclidean 2-norm are the only ones used in this section, so distinguishing subscripts are omitted.

Unitary and Orthogonal Matrices

- A ***unitary matrix*** is defined to be a *complex* matrix $\mathbf{U}_{n\times n}$ whose columns (or rows) constitute an orthonormal basis for $\mathcal{C}^n$.
- An ***orthogonal matrix*** is defined to be a *real* matrix $\mathbf{P}_{n\times n}$ whose columns (or rows) constitute an orthonormal basis for $\Re^n$.

Unitary and orthogonal matrices have some nice features, one of which is the fact that they are easy to invert. To see why, notice that the columns of $\mathbf{U}_{n\times n} = \big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\mathbf{u}_n\big)$ are an orthonormal set if and only if

$$[\mathbf{U}^*\mathbf{U}]_{ij} = \mathbf{u}_i^*\mathbf{u}_j = \begin{cases} 1 & \text{when } i = j, \\ 0 & \text{when } i \neq j, \end{cases} \iff \mathbf{U}^*\mathbf{U} = \mathbf{I} \iff \mathbf{U}^{-1} = \mathbf{U}^*.$$

Notice that because $\mathbf{U}^*\mathbf{U} = \mathbf{I} \iff \mathbf{U}\mathbf{U}^* = \mathbf{I}$, the columns of $\mathbf{U}$ are orthonormal if and only if the rows of $\mathbf{U}$ are orthonormal, and this is why the definitions of unitary and orthogonal matrices can be stated either in terms of orthonormal columns or orthonormal rows.

Another nice feature is that multiplication by a unitary matrix doesn't change the length of a vector. Only the direction can be altered because

$$\|\mathbf{U}\mathbf{x}\|^2 = \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|^2 \quad \forall\ \mathbf{x} \in \mathcal{C}^n. \tag{5.6.1}$$

Conversely, if (5.6.1) holds, then $\mathbf{U}$ must be unitary. To see this, set $\mathbf{x} = \mathbf{e}_i$ in (5.6.1) to observe $\mathbf{u}_i^*\mathbf{u}_i = 1$ for each i, and then set $\mathbf{x} = \mathbf{e}_j + \mathbf{e}_k$ for $j \neq k$ to obtain $0 = \mathbf{u}_j^*\mathbf{u}_k + \mathbf{u}_k^*\mathbf{u}_j = 2\,\mathrm{Re}\,(\mathbf{u}_j^*\mathbf{u}_k)$. By setting $\mathbf{x} = \mathbf{e}_j + \mathrm{i}\mathbf{e}_k$ in (5.6.1) it also follows that $0 = 2\,\mathrm{Im}\,(\mathbf{u}_j^*\mathbf{u}_k)$, so $\mathbf{u}_j^*\mathbf{u}_k = 0$ for each $j \neq k$, and thus (5.6.1) guarantees that $\mathbf{U}$ is unitary.

In the case of orthogonal matrices, everything is real so that $(\star)^*$ can be replaced by $(\star)^T$. Below is a summary of these observations.

Characterizations

- The following statements are equivalent to saying that a complex matrix $\mathbf{U}_{n\times n}$ is unitary.
  - ▷ $\mathbf{U}$ has orthonormal columns.
  - ▷ $\mathbf{U}$ has orthonormal rows.
  - ▷ $\mathbf{U}^{-1} = \mathbf{U}^*$.
  - ▷ $\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \mathcal{C}^{n\times 1}$.
- The following statements are equivalent to saying that a real matrix $\mathbf{P}_{n\times n}$ is orthogonal.
  - ▷ $\mathbf{P}$ has orthonormal columns.
  - ▷ $\mathbf{P}$ has orthonormal rows.
  - ▷ $\mathbf{P}^{-1} = \mathbf{P}^T$.
  - ▷ $\|\mathbf{P}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for every $\mathbf{x} \in \Re^{n\times 1}$.

Example 5.6.1

- The identity matrix $\mathbf{I}$ is an orthogonal matrix.
- All permutation matrices (products of elementary interchange matrices) are orthogonal—recall Exercise 3.9.4.
- The matrix
$$\mathbf{P} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{3} & -1/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{3} & -1/\sqrt{6} \\ 0 & 1/\sqrt{3} & 2/\sqrt{6} \end{pmatrix}$$
is an orthogonal matrix because $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ or, equivalently, because the columns (and rows) constitute an orthonormal set.
- The matrix $\mathbf{U} = \frac{1}{2}\begin{pmatrix} 1+\mathrm{i} & -1+\mathrm{i} \\ 1+\mathrm{i} & 1-\mathrm{i} \end{pmatrix}$ is unitary because $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$ or, equivalently, because the columns (and rows) are an orthonormal set.
- An orthogonal matrix can be considered to be unitary, but a unitary matrix is generally not orthogonal.

In general, a linear operator $\mathbf{T}$ on a vector space $\mathcal{V}$ with the property that $\|\mathbf{T}\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathcal{V}$ is called an ***isometry*** on $\mathcal{V}$. The isometries on $\Re^n$ are precisely the orthogonal matrices, and the isometries on $\mathcal{C}^n$ are the unitary matrices. The term "isometry" has an advantage in that it can be used to treat the real and complex cases simultaneously, but for clarity we will often revert back to the more cumbersome "orthogonal" and "unitary" terminology.

The geometrical concepts of projection, reflection, and rotation are among the most fundamental of all linear transformations in $\Re^2$ and $\Re^3$ (see Example 4.7.1 for three simple examples), so pursuing these ideas in higher dimensions is only natural. The reflector and rotator given in Example 4.7.1 are isometries (because they preserve length), but the projector is not. We are about to see that the same is true in more general settings.

Elementary Orthogonal Projectors

For a vector $\mathbf{u} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{u}\| = 1$, a matrix of the form

$$\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^* \tag{5.6.2}$$

is called an ***elementary orthogonal projector.*** More general projectors are discussed on pp. 386 and 429.

To understand the nature of elementary projectors consider the situation in $\Re^3$. Suppose that $\|\mathbf{u}_{3\times 1}\| = 1$, and let $\mathbf{u}^\perp$ denote the space (the plane through the origin) consisting of all vectors that are perpendicular to $\mathbf{u}$—we call $\mathbf{u}^\perp$ the ***orthogonal complement*** of $\mathbf{u}$ (a more general definition appears on p. 403). The matrix $\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto $\mathbf{u}^\perp$ in the sense that $\mathbf{Q}$ maps each $\mathbf{x} \in \Re^{3\times 1}$ to its orthogonal projection in $\mathbf{u}^\perp$ as shown in Figure 5.6.1.

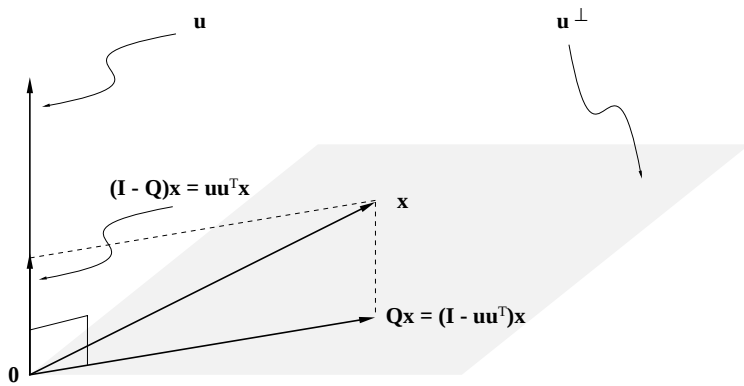


FIGURE 5.6.1

To see this, observe that each $\mathbf{x}$ can be resolved into two components

$$\mathbf{x} = (\mathbf{I} - \mathbf{Q})\mathbf{x} + \mathbf{Q}\mathbf{x}, \quad \text{where} \quad (\mathbf{I} - \mathbf{Q})\mathbf{x} \perp \mathbf{Q}\mathbf{x}.$$

The vector $(\mathbf{I} - \mathbf{Q})\mathbf{x} = \mathbf{u}(\mathbf{u}^T\mathbf{x})$ is on the line determined by $\mathbf{u}$, and $\mathbf{Q}\mathbf{x}$ is in the plane $\mathbf{u}^\perp$ because $\mathbf{u}^T\mathbf{Q}\mathbf{x} = \mathbf{0}$.

The situation is exactly as depicted in Figure 5.6.1. Notice that $(\mathbf{I}-\mathbf{Q})\mathbf{x} = \mathbf{u}\mathbf{u}^T\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the line determined by $\mathbf{u}$ and $\|\mathbf{u}\mathbf{u}^T\mathbf{x}\| = |\mathbf{u}^T\mathbf{x}|$. This provides a nice interpretation of the magnitude of the standard inner product. Below is a summary.

Geometry of Elementary Projectors

For vectors $\mathbf{u}, \mathbf{x} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{u}\| = 1$,

- $(\mathbf{I} - \mathbf{u}\mathbf{u}^*)\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the orthogonal complement $\mathbf{u}^\perp$, the space of all vectors orthogonal to $\mathbf{u}$; (5.6.3)
- $\mathbf{u}\mathbf{u}^*\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto the one-dimensional space $span\,\{\mathbf{u}\}$; (5.6.4)
- $|\mathbf{u}^*\mathbf{x}|$ represents the length of the orthogonal projection of $\mathbf{x}$ onto the one-dimensional space $span\,\{\mathbf{u}\}$. (5.6.5)

In passing, note that elementary projectors are never isometries—they can't be because they are not unitary matrices in the complex case and not orthogonal matrices in the real case. Furthermore, isometries are nonsingular but elementary projectors are singular.

Example 5.6.2

Problem: Determine the orthogonal projection of $\mathbf{x}$ onto $span\,\{\mathbf{u}\}$, and then find the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^\perp$ for $\mathbf{x} = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$ and $\mathbf{u} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}$.

Solution: We cannot apply (5.6.3) and (5.6.4) directly because $\|\mathbf{u}\| \neq 1$, but this is not a problem because

$$\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|} \right\| = 1, \qquad span\,\{\mathbf{u}\} = span\left\{ \frac{\mathbf{u}}{\|\mathbf{u}\|} \right\}, \qquad \text{and} \qquad \mathbf{u}^\perp = \left(\frac{\mathbf{u}}{\|\mathbf{u}\|} \right)^\perp.$$

Consequently, the orthogonal projection of $\mathbf{x}$ onto $span\,\{\mathbf{u}\}$ is given by

$$\left(\frac{\mathbf{u}}{\|\mathbf{u}\|} \right) \left(\frac{\mathbf{u}}{\|\mathbf{u}\|} \right)^T \mathbf{x} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{x} = \frac{1}{2} \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix},$$

and the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^\perp$ is

$$\left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} \right) \mathbf{x} = \mathbf{x} - \frac{\mathbf{u}\mathbf{u}^T\mathbf{x}}{\mathbf{u}^T\mathbf{u}} = \frac{1}{2} \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}.$$

There is nothing special about the numbers in this example. For every nonzero vector $\mathbf{u} \in \mathcal{C}^{n\times 1}$, the orthogonal projectors onto $span\,\{\mathbf{u}\}$ and $\mathbf{u}^{\perp}$ are

$$\mathbf{P}_{\mathbf{u}} = \frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}} \qquad \text{and} \qquad \mathbf{P}_{\mathbf{u}^{\perp}} = \mathbf{I} - \frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}. \tag{5.6.6}$$

Elementary Reflectors

For $\mathbf{u}_{n\times 1} \neq \mathbf{0}$, the ***elementary reflector*** about $\mathbf{u}^{\perp}$ is defined to be

$$\mathbf{R} = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}} \tag{5.6.7}$$

or, equivalently,

$$\mathbf{R} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^* \quad \text{when} \quad \|\mathbf{u}\| = 1. \tag{5.6.8}$$

Elementary reflectors are also called ***Householder transformations,***[46] and they are analogous to the simple reflector given in Example 4.7.1. To understand why, suppose $\mathbf{u} \in \Re^{3\times 1}$ and $\|\mathbf{u}\| = 1$ so that $\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto the plane $\mathbf{u}^{\perp}$. For each $\mathbf{x} \in \Re^{3\times 1}$, $\mathbf{Q}\mathbf{x}$ is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^{\perp}$ as shown in Figure 5.6.1. To locate $\mathbf{R}\mathbf{x} = (\mathbf{I} - 2\mathbf{u}\mathbf{u}^T)\mathbf{x}$, notice that $\mathbf{Q}(\mathbf{R}\mathbf{x}) = \mathbf{Q}\mathbf{x}$. In other words, $\mathbf{Q}\mathbf{x}$ is simultaneously the orthogonal projection of $\mathbf{x}$ onto $\mathbf{u}^{\perp}$ as well as the orthogonal projection of $\mathbf{R}\mathbf{x}$ onto $\mathbf{u}^{\perp}$. This together with $\|\mathbf{x} - \mathbf{Q}\mathbf{x}\| = |\mathbf{u}^T\mathbf{x}| = \|\mathbf{Q}\mathbf{x} - \mathbf{R}\mathbf{x}\|$ implies that $\mathbf{R}\mathbf{x}$ *is the reflection of* $\mathbf{x}$ *about the plane* $\mathbf{u}^{\perp}$, exactly as depicted in Figure 5.6.2. (Reflections about more general subspaces are examined in Exercise 5.13.21.)

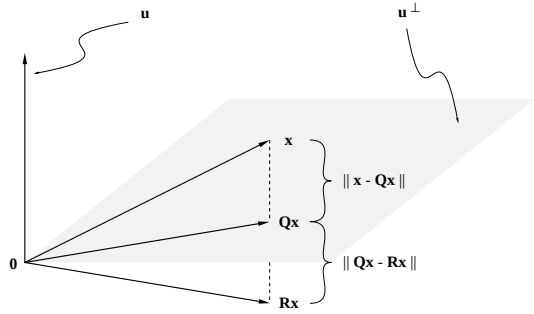


FIGURE 5.6.2

[46] Alston Scott Householder (1904–1993) was one of the first people to appreciate and promote the use of elementary reflectors for numerical applications. Although his 1937 Ph.D. dissertation at University of Chicago concerned the calculus of variations, Householder's passion was mathematical biology, and this was the thrust of his career until it was derailed by the war effort in 1944. Householder joined the Mathematics Division of Oak Ridge National Laboratory in 1946 and became its director in 1948. He stayed at Oak Ridge for the remainder of his career, and he became a leading figure in numerical analysis and matrix computations. Like his counterpart J. Wallace Givens (p. 333) at the Argonne National Laboratory, Householder was one of the early presidents of SIAM.

Properties of Elementary Reflectors

- All elementary reflectors $\mathbf{R}$ are unitary, hermitian, and involutory ($\mathbf{R}^2 = \mathbf{I}$). That is,
$$\mathbf{R} = \mathbf{R}^* = \mathbf{R}^{-1}. \tag{5.6.9}$$
- If $\mathbf{x}_{n\times 1}$ is a vector whose first entry is $x_1 \neq 0$, and if
$$\mathbf{u} = \mathbf{x} \pm \mu \|\mathbf{x}\| \mathbf{e}_1, \quad \text{where} \quad \mu = \begin{cases} 1 & \text{if } x_1 \text{ is real,} \\ x_1/|x_1| & \text{if } x_1 \text{ is not real,} \end{cases} \tag{5.6.10}$$
is used to build the elementary reflector $\mathbf{R}$ in (5.6.7), then
$$\mathbf{R}\mathbf{x} = \mp\mu \|\mathbf{x}\| \mathbf{e}_1. \tag{5.6.11}$$
In other words, this $\mathbf{R}$ "reflects" $\mathbf{x}$ onto the first coordinate axis.
Computational Note: To avoid cancellation when using floating-point arithmetic for real matrices, set $\mathbf{u} = \mathbf{x} + \text{sign}(x_1) \|\mathbf{x}\| \mathbf{e}_1$.

Proof of (5.6.9). It is clear that $\mathbf{R} = \mathbf{R}^*$, and the fact that $\mathbf{R} = \mathbf{R}^{-1}$ is established simply by verifying that $\mathbf{R}^2 = \mathbf{I}$.

Proof of (5.6.10). Observe that $\mathbf{R} = \mathbf{I} - 2\hat{\mathbf{u}}\hat{\mathbf{u}}^*$, where $\hat{\mathbf{u}} = \mathbf{u}/\|\mathbf{u}\|$.

Proof of (5.6.11). Write $\mathbf{R}\mathbf{x} = \mathbf{x} - 2\mathbf{u}\mathbf{u}^*\mathbf{x}/\mathbf{u}^*\mathbf{u} = \mathbf{x} - (2\mathbf{u}^*\mathbf{x}/\mathbf{u}^*\mathbf{u})\mathbf{u}$ and verify that $2\mathbf{u}^*\mathbf{x} = \mathbf{u}^*\mathbf{u}$ to conclude $\mathbf{R}\mathbf{x} = \mathbf{x} - \mathbf{u} = \mp\mu \|\mathbf{x}\| \mathbf{e}_1$. ∎

Example 5.6.3

Problem: Given $\mathbf{x} \in \mathcal{C}^{n\times 1}$ such that $\|\mathbf{x}\| = 1$, construct an orthonormal basis for $\mathcal{C}^n$ that contains $\mathbf{x}$.

Solution: An efficient solution is to build a unitary matrix that contains $\mathbf{x}$ as its first column. Set $\mathbf{u} = \mathbf{x} \pm \mu\mathbf{e}_1$ in $\mathbf{R} = \mathbf{I} - 2(\mathbf{u}\mathbf{u}^*/\mathbf{u}^*\mathbf{u})$ and notice that (5.6.11) guarantees $\mathbf{R}\mathbf{x} = \mp\mu\mathbf{e}_1$, so multiplication on the left by $\mathbf{R}$ (remembering that $\mathbf{R}^2 = \mathbf{I}$) produces $\mathbf{x} = \mp\mu\mathbf{R}\mathbf{e}_1 = [\mp\mu\mathbf{R}]_{*1}$. Since $|\mp\mu| = 1$, $\mathbf{U} = \mp\mu\mathbf{R}$ is a unitary matrix with $\mathbf{U}_{*1} = \mathbf{x}$, so the columns of $\mathbf{U}$ provide the desired orthonormal basis. For example, to construct an orthonormal basis for $\Re^4$ that includes $\mathbf{x} = (1/3)\begin{pmatrix} -1 & 2 & 0 & -2 \end{pmatrix}^T$, set

$$\mathbf{u} = \mathbf{x} - \mathbf{e}_1 = \frac{1}{3}\begin{pmatrix} -4 \\ 2 \\ 0 \\ -2 \end{pmatrix} \text{ and compute } \mathbf{R} = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{3}\begin{pmatrix} -1 & 2 & 0 & -2 \\ 2 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \\ -2 & 1 & 0 & 2 \end{pmatrix}.$$

The columns of $\mathbf{R}$ do the job.

Now consider rotation, and begin with a basic problem in $\Re^2$. If a nonzero vector $\mathbf{u} = (u_1, u_2)$ is rotated counterclockwise through an angle θ to produce $\mathbf{v} = (v_1, v_2)$, how are the coordinates of $\mathbf{v}$ related to the coordinates of $\mathbf{u}$? To answer this question, refer to Figure 5.6.3, and use the fact that $\|\mathbf{u}\| = \nu = \|\mathbf{v}\|$ (rotation is an isometry) together with some elementary trigonometry to obtain

$$\begin{aligned} v_1 &= \nu\cos(\phi+\theta) = \nu(\cos\theta\cos\phi - \sin\theta\sin\phi), \\ v_2 &= \nu\sin(\phi+\theta) = \nu(\sin\theta\cos\phi + \cos\theta\sin\phi). \end{aligned} \tag{5.6.12}$$

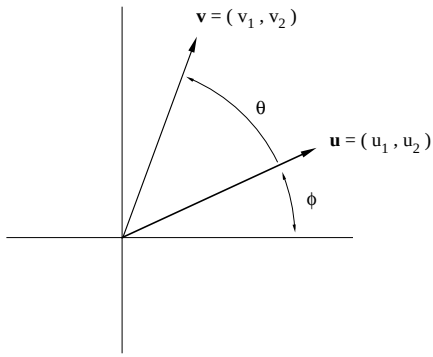


FIGURE 5.6.3

Substituting $\cos\phi = u_1/\nu$ and $\sin\phi = u_2/\nu$ into (5.6.12) yields

$$\begin{aligned} v_1 &= (\cos\theta)u_1 - (\sin\theta)u_2, \\ v_2 &= (\sin\theta)u_1 + (\cos\theta)u_2, \end{aligned} \quad \text{or} \quad \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{5.6.13}$$

In other words, $\mathbf{v} = \mathbf{P}\mathbf{u}$, where $\mathbf{P}$ is the ***rotator*** (rotation operator)

$$\mathbf{P} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{5.6.14}$$

Notice that $\mathbf{P}$ is an orthogonal matrix because $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. This means that if $\mathbf{v} = \mathbf{P}\mathbf{u}$, then $\mathbf{u} = \mathbf{P}^T\mathbf{v}$, and hence $\mathbf{P}^T$ is also a rotator, but in the opposite direction of that associated with $\mathbf{P}$. That is, $\mathbf{P}^T$ is the rotator associated with the angle $-\theta$. This is confirmed by the fact that if θ is replaced by $-\theta$ in (5.6.14), then $\mathbf{P}^T$ is produced.

Rotating vectors in $\Re^3$ around any one of the coordinate axes is similar. For example, consider rotation around the z-axis. Suppose that $\mathbf{v} = (v_1, v_2, v_3)$ is obtained by rotating $\mathbf{u} = (u_1, u_2, u_3)$ counterclockwise[47] through an angle θ around the z-axis. Just as before, the goal is to determine the relationship between the coordinates of $\mathbf{u}$ and $\mathbf{v}$. Since we are rotating around the z-axis,

[47] This is from the perspective of looking down the z-axis onto the xy-plane.

it is evident (see Figure 5.6.4) that the third coordinates are unaffected—i.e., $v_3 = u_3$. To see how the xy-coordinates of $\mathbf{u}$ and $\mathbf{v}$ are related, consider the orthogonal projections

$$\mathbf{u}_p = (u_1,\, u_2,\, 0) \quad \text{and} \quad \mathbf{v}_p = (v_1,\, v_2,\, 0)$$

of $\mathbf{u}$ and $\mathbf{v}$ onto the xy-plane.

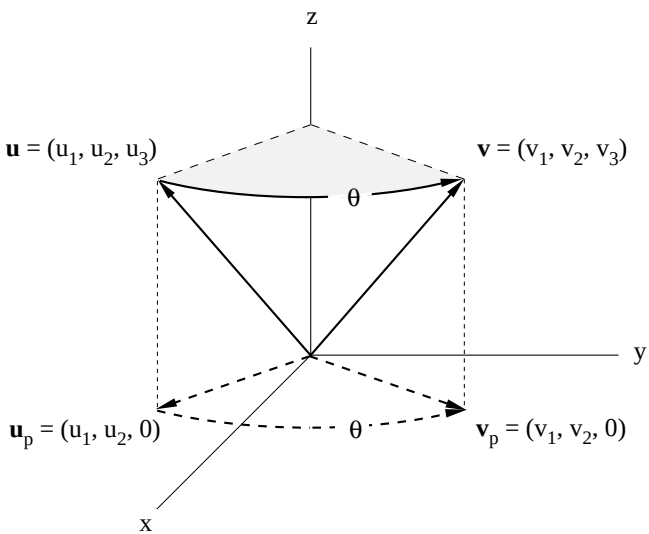


FIGURE 5.6.4

It's apparent from Figure 5.6.4 that the problem has been reduced to rotation in the xy-plane, and we already know how to do this. Combining (5.6.13) with the fact that $v_3 = u_3$ produces the equation

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix},$$

so

$$\mathbf{P}_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is the matrix that rotates vectors in $\Re^3$ counterclockwise around the z-axis through an angle θ. It is easy to verify that $\mathbf{P}_z$ is an orthogonal matrix and that $\mathbf{P}_z^{-1} = \mathbf{P}_z^T$ rotates vectors *clockwise* around the z-axis.

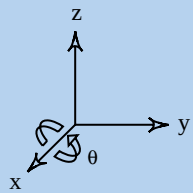
By using similar techniques, it is possible to derive orthogonal matrices that rotate vectors around the x-axis or around the y-axis. Below is a summary of these rotations in $\Re^3$.

Rotations in $\mathbf{R}^3$

A vector $\mathbf{u} \in \Re^3$ can be rotated counterclockwise through an angle θ around a coordinate axis by means of a multiplication $\mathbf{P}_\star \mathbf{u}$ in which $\mathbf{P}_\star$ is an appropriate orthogonal matrix as described below.

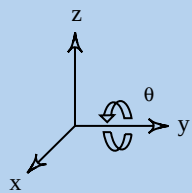
Rotation around the x-Axis

$$\mathbf{P}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix}$$



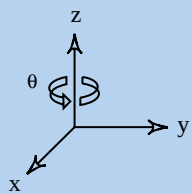
Rotation around the y-Axis

$$\mathbf{P}_y = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}$$



Rotation around the z-Axis

$$\mathbf{P}_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$



Note: The minus sign appears above the diagonal in $\mathbf{P}_x$ and $\mathbf{P}_z$, but below the diagonal in $\mathbf{P}_y$. This is not a mistake—it's due to the orientation of the positive x-axis with respect to the yz-plane.

Example 5.6.4

3-D Rotational Coordinates. Suppose that three counterclockwise rotations are performed on the three-dimensional solid shown in Figure 5.6.5. First rotate the solid in View (a) $90°$ around the x-axis to obtain the orientation shown in View (b). Then rotate View (b) $45°$ around the y-axis to produce View (c) and, finally, rotate View (c) $60°$ around the z-axis to end up with View (d). You can follow the process by watching how the notch, the vertex $\mathbf{v}$, and the lighter shaded face move.

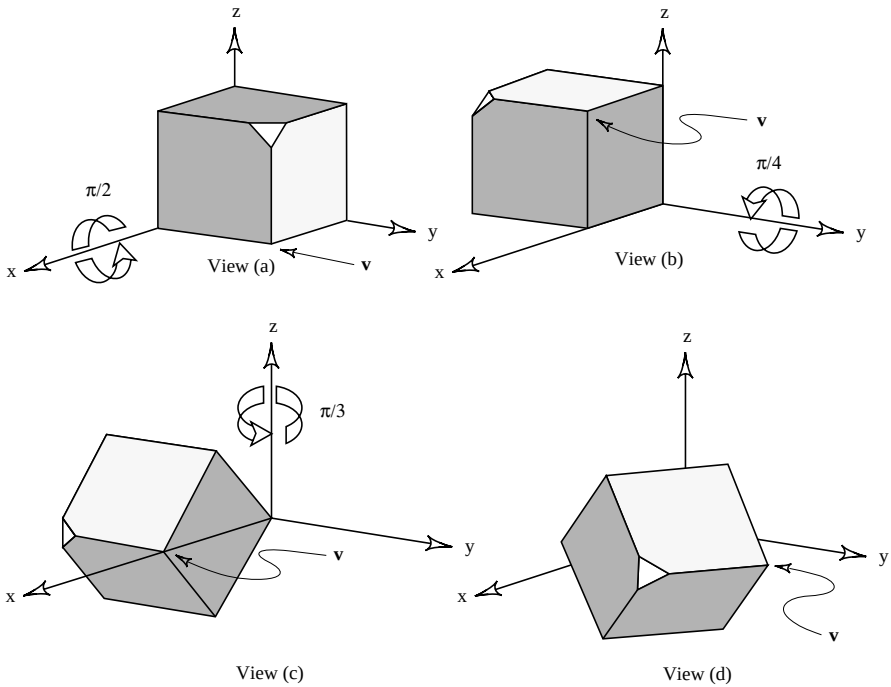


FIGURE 5.6.5

Problem: If the coordinates of each vertex in View (a) are specified, what are the coordinates of each vertex in View (d)?

Solution: If $\mathbf{P}_x$ is the rotator that maps points in View (a) to corresponding points in View (b), and if $\mathbf{P}_y$ and $\mathbf{P}_z$ are the respective rotators carrying View (b) to View (c) and View (c) to View (d), then

$$\mathbf{P}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \ \mathbf{P}_y = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 1 \\ 0 & \sqrt{2} & 0 \\ -1 & 0 & 1 \end{pmatrix}, \ \mathbf{P}_z = \begin{pmatrix} 1/2 & -\sqrt{3}/2 & 0 \\ \sqrt{3}/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

so

$$\mathbf{P} = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x = \frac{1}{2\sqrt{2}} \begin{pmatrix} 1 & 1 & \sqrt{6} \\ \sqrt{3} & \sqrt{3} & -\sqrt{2} \\ -2 & 2 & 0 \end{pmatrix} \tag{5.6.15}$$

is the orthogonal matrix that maps points in View (a) to their corresponding images in View (d). For example, focus on the vertex labeled $\mathbf{v}$ in View (a), and let $\mathbf{v}_a$, $\mathbf{v}_b$, $\mathbf{v}_c$, and $\mathbf{v}_d$ denote its respective coordinates in Views (a), (b), (c), and (d). If $\mathbf{v}_a = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}^T$, then $\mathbf{v}_b = \mathbf{P}_x\mathbf{v}_a = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^T$,

$$\mathbf{v}_c = \mathbf{P}_y\mathbf{v}_b = \mathbf{P}_y\mathbf{P}_x\mathbf{v}_a = \begin{pmatrix} \sqrt{2} \\ 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{v}_d = \mathbf{P}_z\mathbf{v}_c = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x\mathbf{v}_a = \begin{pmatrix} \sqrt{2}/2 \\ \sqrt{6}/2 \\ 0 \end{pmatrix}.$$

More generally, if the coordinates of each of the ten vertices in View (a) are placed as columns in a ***vertex matrix,***

$$\mathbf{V}_a = \begin{array}{c} \begin{array}{cccc} \mathbf{v}_1 & \mathbf{v}_2 & & \mathbf{v}_{10} \\ \downarrow & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} x_1 & x_2 & \cdots & x_{10} \\ y_1 & y_2 & \cdots & y_{10} \\ z_1 & z_2 & \cdots & z_{10} \end{pmatrix} \end{array}, \text{ then } \quad \mathbf{V}_d = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x\mathbf{V}_a = \begin{array}{c} \begin{array}{cccc} \hat{\mathbf{v}}_1 & \hat{\mathbf{v}}_2 & & \hat{\mathbf{v}}_{10} \\ \downarrow & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} \hat{x}_1 & \hat{x}_2 & \cdots & \hat{x}_{10} \\ \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_{10} \\ \hat{z}_1 & \hat{z}_2 & \cdots & \hat{z}_{10} \end{pmatrix} \end{array}$$

is the vertex matrix for the orientation shown in View (d). The polytope in View (d) is drawn by identifying pairs of vertices $(\mathbf{v}_i, \mathbf{v}_j)$ in $\mathbf{V}_a$ that have an edge between them, and by drawing an edge between the corresponding vertices $(\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_j)$ in $\mathbf{V}_d$.

Example 5.6.5

3-D Computer Graphics. Consider the problem of displaying and manipulating views of a three-dimensional solid on a two-dimensional computer display monitor. One simple technique is to use a *wire-frame representation* of the solid consisting of a mesh of points (vertices) on the solid's surface connected by straight line segments (edges). Once these vertices and edges have been defined, the resulting polytope can be oriented in any desired manner as described in Example 5.6.4, so all that remains are the following problems.

Problem: How should the vertices and edges of a three-dimensional polytope be plotted on a two-dimensional computer monitor?

Solution: Assume that the screen represents the yz-plane, and suppose the x-axis is orthogonal to the screen so that it points toward the viewer's eye as shown in Figure 5.6.6.

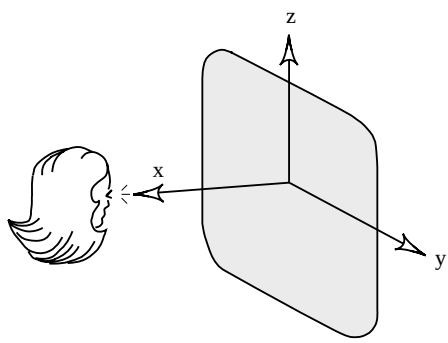


FIGURE 5.6.6

A solid in the xyz-coordinate system appears to the viewer as the orthogonal projection of the solid onto the yz-plane, and the projection of a polytope is easy to draw. Just set the x-coordinate of each vertex to 0 (i.e., ignore the x-coordinates), plot the (y, z)-coordinates on the yz-plane (the screen), and

draw edges between appropriate vertices. For example, suppose that the vertices of the polytope in Figure 5.6.5 are numbered as indicated below in Figure 5.6.7,

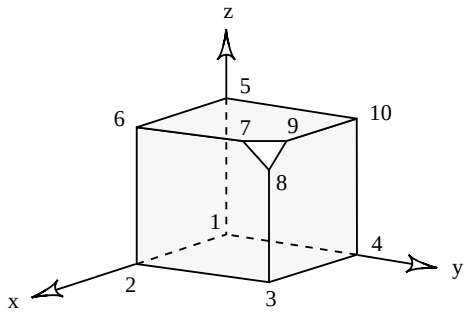


FIGURE 5.6.7

and suppose that the associated vertex matrix is

$$\mathbf{V} = \begin{array}{c} \\ x \\ y \\ z \end{array} \begin{array}{c} \begin{array}{cccccccccc} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5 & \mathbf{v}_6 & \mathbf{v}_7 & \mathbf{v}_8 & \mathbf{v}_9 & \mathbf{v}_{10} \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & .8 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & .8 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & .8 & 1 & 1 \end{pmatrix} \end{array}.$$

There are 15 edges, and they can be recorded in an ***edge matrix***

$$\mathbf{E} = \begin{array}{c} \begin{array}{ccccccccccccccc} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 & \mathbf{e}_5 & \mathbf{e}_6 & \mathbf{e}_7 & \mathbf{e}_8 & \mathbf{e}_9 & \mathbf{e}_{10} & \mathbf{e}_{11} & \mathbf{e}_{12} & \mathbf{e}_{13} & \mathbf{e}_{14} & \mathbf{e}_{15} \end{array} \\ \begin{pmatrix} 1 & 2 & 3 & 4 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 7 & 8 & 9 & 10 \\ 2 & 3 & 4 & 1 & 5 & 6 & 8 & 10 & 6 & 7 & 8 & 9 & 9 & 10 & 5 \end{pmatrix} \end{array}$$

in which the kth column represents an edge between the indicated pair of vertices. To display the image of the polytope in Figure 5.6.7 on a monitor, (i) drop the first row from $\mathbf{V}$, (ii) plot the remaining yz-coordinates on the screen, (iii) draw edges between appropriate vertices as dictated by the information in the edge matrix $\mathbf{E}$. To display the image of the polytope after it has been rotated counterclockwise around the x-, y-, and z-axes by $90°$, $45°$, and $60°$, respectively, use the orthogonal matrix $\mathbf{P} = \mathbf{P}_z\mathbf{P}_y\mathbf{P}_x$ determined in (5.6.15) and compute the product

$$\mathbf{PV} = \begin{pmatrix} 0 & .354 & .707 & .354 & .866 & 1.22 & 1.5 & 1.4 & 1.5 & 1.22 \\ 0 & .612 & 1.22 & .612 & -.5 & .112 & .602 & .825 & .602 & .112 \\ 0 & -.707 & 0 & .707 & 0 & -.707 & -.141 & 0 & .141 & .707 \end{pmatrix}.$$

Now proceed as before—(i) ignore the first row of $\mathbf{PV}$, (ii) plot the points in the second and third row of $\mathbf{PV}$ as yz-coordinates on the monitor, (iii) draw edges between appropriate vertices as indicated by the edge matrix $\mathbf{E}$.

Problem: In addition to rotation, how can a polytope (or its image on a computer monitor) be translated?

Solution: Translation of a polytope to a different point in space is accomplished by adding a constant to each of its coordinates. For example, to translate the polytope shown in Figure 5.6.7 to the location where vertex 1 is at $\mathbf{p}^T = (x_0, y_0, z_0)$ instead of at the origin, just add $\mathbf{p}$ to every point. In particular, if $\mathbf{e}$ is the column of 1's, the translated vertex matrix is

$$\mathbf{V}_{trans} = \mathbf{V}_{orig} + \begin{pmatrix} x_0 & x_0 & \cdots & x_0 \\ y_0 & y_0 & \cdots & y_0 \\ z_0 & z_0 & \cdots & z_0 \end{pmatrix} = \mathbf{V}_{orig} + \mathbf{p}\mathbf{e}^T \quad \text{(a rank-1 update).}$$

Of course, the edge matrix is not affected by translation.

Problem: How can a polytope (or its image on a computer monitor) be scaled?

Solution: Simply multiply every coordinate by the desired scaling factor. For example, to scale an image by a factor α, form the scaled vertex matrix

$$\mathbf{V}_{scaled} = \alpha \mathbf{V}_{orig},$$

and then connect the scaled vertices with appropriate edges as dictated by the edge matrix $\mathbf{E}$.

Problem: How can the faces of a polytope that are hidden from the viewer's perspective be detected so that they can be omitted from the drawing on the screen?

Solution: A complete discussion of this tricky problem would carry us too far astray, but one clever solution relying on the cross product of vectors in $\Re^3$ is presented in Exercise 5.6.21 for the case of *convex* polytopes.

Rotations in higher dimensions are straightforward generalizations of rotations in $\Re^3$. Recall from p. 328 that rotation around any particular axis in $\Re^3$ amounts to rotation in the complementary plane, and the associated 3×3 rotator is constructed by embedding a 2×2 rotator in the appropriate position in a 3×3 identity matrix. For example, rotation around the y-axis is rotation in the xz-plane, and the corresponding rotator is produced by embedding

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

in the "xz-position" of $\mathbf{I}_{3\times 3}$ to form

$$\mathbf{P}_y = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}.$$

These observations directly extend to higher dimensions.

Plane Rotations

Orthogonal matrices of the form

$$\mathbf{P}_{ij} = \begin{array}{c} \begin{array}{cccccccc} & & \text{col } i & & & \text{col } j & & \end{array} \\ \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & c & & & s & & \\ & & & 1 & & & & \\ & & & & \ddots & & & \\ & & -s & & & c & & \\ & & & & & & 1 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \longleftarrow \text{row } i \\ \\ \\ \longleftarrow \text{row } j \\ \\ \\ \\ \end{array}$$

in which $c^2 + s^2 = 1$ are called ***plane rotation matrices*** because they perform a rotation in the (i,j)-plane of $\Re^n$. The entries c and s are meant to suggest cosine and sine, respectively, but designating a rotation angle θ as is done in $\Re^2$ and $\Re^3$ is not useful in higher dimensions.

Plane rotations matrices $\mathbf{P}_{ij}$ are also called ***Givens***[48] ***rotations***. Applying $\mathbf{P}_{ij}$ to $\mathbf{0} \neq \mathbf{x} \in \Re^n$ rotates the (i,j)-coordinates of $\mathbf{x}$ in the sense that

$$\mathbf{P}_{ij}\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ cx_i + sx_j \\ \vdots \\ -sx_i + cx_j \\ \vdots \\ x_n \end{pmatrix} \begin{array}{l} \\ \\ \longleftarrow i \\ \\ \longleftarrow j \\ \\ \\ \end{array}.$$

If x_i and x_j are not both zero, and if we set

$$c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} \quad \text{and} \quad s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}}, \tag{5.6.16}$$

[48] J. Wallace Givens, Jr. (1910–1993) pioneered the use of plane rotations in the early days of automatic matrix computations. Givens graduated from Lynchburg College in 1928, and he completed his Ph.D. at Princeton University in 1936. After spending three years at the Institute for Advanced Study in Princeton as an assistant of O. Veblen, Givens accepted an appointment at Cornell University but later moved to Northwestern University. In addition to his academic career, Givens was the Director of the Applied Mathematics Division at Argonne National Laboratory and, like his counterpart A. S. Householder (p. 324) at Oak Ridge National Laboratory, Givens served as an early president of SIAM.

then

$$\mathbf{P}_{ij}\mathbf{x} = \begin{array}{cl} \begin{pmatrix} x_1 \\ \vdots \\ \sqrt{x_i^2 + x_j^2} \\ \vdots \\ 0 \\ \vdots \\ x_n \end{pmatrix} & \begin{array}{l} \\ \\ \longleftarrow i \\ \\ \longleftarrow j \\ \\ \\ \end{array} \end{array}.$$

This means that we can selectively annihilate any component—the jth in this case—by a rotation in the (i,j)-plane without affecting any entry except x_i and x_j. Consequently, plane rotations can be applied to annihilate *all* components below any particular "pivot." For example, to annihilate all entries below the first position in $\mathbf{x}$, apply a sequence of plane rotations as follows:

$$\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \sqrt{x_1^2+x_2^2} \\ 0 \\ x_3 \\ x_4 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{P}_{13}\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \sqrt{x_1^2+x_2^2+x_3^2} \\ 0 \\ 0 \\ x_4 \\ \vdots \\ x_n \end{pmatrix}, \quad \ldots, \quad \mathbf{P}_{1n}\cdots\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The product of plane rotations is generally not another plane rotation, but such a product is always an orthogonal matrix, and hence it is an isometry. If we are willing to interpret "rotation in $\Re^n$" as a sequence of plane rotations, then we can say that it is always possible to "rotate" each nonzero vector onto the first coordinate axis. Recall from (5.6.11) that we can also do this with a reflection. More generally, the following statement is true.

Rotations in $\Re^n$

Every nonzero vector $\mathbf{x} \in \Re^n$ can be rotated to the ith coordinate axis by a sequence of $n-1$ plane rotations. In other words, there is an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{x} = \|\mathbf{x}\|\,\mathbf{e}_i, \tag{5.6.17}$$

where $\mathbf{P}$ has the form

$$\mathbf{P} = \mathbf{P}_{in}\cdots\mathbf{P}_{i,i+1}\mathbf{P}_{i,i-1}\cdots\mathbf{P}_{i1}.$$

Example 5.6.6

Problem: If $\mathbf{x} \in \Re^n$ is a vector such that $\|\mathbf{x}\| = 1$, explain how to use plane rotations to construct an orthonormal basis for $\Re^n$ that contains $\mathbf{x}$.

Solution: This is almost the same problem as that posed in Example 5.6.3, and, as explained there, the goal is to construct an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}_{*1} = \mathbf{x}$. But this time we need to use plane rotations rather than an elementary reflector. Equation (5.6.17) asserts that we can build an orthogonal matrix from a sequence of plane rotations $\mathbf{P} = \mathbf{P}_{1n} \cdots \mathbf{P}_{13}\mathbf{P}_{12}$ such that $\mathbf{P}\mathbf{x} = \mathbf{e}_1$. Thus $\mathbf{x} = \mathbf{P}^T\mathbf{e}_1 = \mathbf{P}_{*1}^T$, and hence the columns of $\mathbf{Q} = \mathbf{P}^T$ serve the purpose. For example, to extend

$$\mathbf{x} = \frac{1}{3}\begin{pmatrix} -1 \\ 2 \\ 0 \\ -2 \end{pmatrix}$$

to an orthonormal basis for $\Re^4$, sequentially annihilate the second and fourth components of $\mathbf{x}$ by using (5.6.16) to construct the following plane rotations:

$$\mathbf{P}_{12}\mathbf{x} = \begin{pmatrix} -1/\sqrt{5} & 2/\sqrt{5} & 0 & 0 \\ -2/\sqrt{5} & -1/\sqrt{5} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \frac{1}{3}\begin{pmatrix} -1 \\ 2 \\ 0 \\ -2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} \sqrt{5} \\ 0 \\ 0 \\ -2 \end{pmatrix},$$

$$\mathbf{P}_{14}\Big(\mathbf{P}_{12}\mathbf{x}\Big) = \begin{pmatrix} \sqrt{5}/3 & 0 & 0 & -2/3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2/3 & 0 & 0 & \sqrt{5}/3 \end{pmatrix} \frac{1}{3}\begin{pmatrix} \sqrt{5} \\ 0 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Therefore, the columns of

$$\mathbf{Q} = (\mathbf{P}_{14}\mathbf{P}_{12})^T = \mathbf{P}_{12}^T\mathbf{P}_{14}^T = \begin{pmatrix} -1/3 & -2/\sqrt{5} & 0 & -2/3\sqrt{5} \\ 2/3 & -1/\sqrt{5} & 0 & 4/3\sqrt{5} \\ 0 & 0 & 1 & 0 \\ -2/3 & 0 & 0 & \sqrt{5}/3 \end{pmatrix}$$

are an orthonormal set containing the specified vector $\mathbf{x}$.

Exercises for section 5.6

5.6.1. Determine which of the following matrices are isometries.

(a) $\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix}$. (b) $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$.

(c) $\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$. (d) $\begin{pmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{i\theta_n} \end{pmatrix}$.

5.6.2. Is $\begin{pmatrix} \dfrac{1+i}{\sqrt{3}} & \dfrac{1+i}{\sqrt{6}} \\ \dfrac{i}{\sqrt{3}} & \dfrac{-2\,i}{\sqrt{6}} \end{pmatrix}$ a unitary matrix?

5.6.3. (a) How many 3×3 matrices are both diagonal and orthogonal?
(b) How many $n \times n$ matrices are both diagonal and orthogonal?
(c) How many $n \times n$ matrices are both diagonal and unitary?

5.6.4. (a) Under what conditions on the real numbers α and β will

$$\mathbf{P} = \begin{pmatrix} \alpha+\beta & \beta-\alpha \\ \alpha-\beta & \beta+\alpha \end{pmatrix}$$

be an orthogonal matrix?
(b) Under what conditions on the real numbers α and β will

$$\mathbf{U} = \begin{pmatrix} 0 & \alpha & 0 & \mathrm{i}\beta \\ \alpha & 0 & \mathrm{i}\beta & 0 \\ 0 & \mathrm{i}\beta & 0 & \alpha \\ \mathrm{i}\beta & 0 & \alpha & 0 \end{pmatrix}$$

be a unitary matrix?

5.6.5. Let $\mathbf{U}$ and $\mathbf{V}$ be two $n \times n$ unitary (orthogonal) matrices.
(a) Explain why the product $\mathbf{UV}$ must be unitary (orthogonal).
(b) Explain why the sum $\mathbf{U}+\mathbf{V}$ need not be unitary (orthogonal).
(c) Explain why $\begin{pmatrix} \mathbf{U}_{n\times n} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{m\times m} \end{pmatrix}$ must be unitary (orthogonal).

5.6.6. Cayley Transformation. Prove, as Cayley did in 1846, that if $\mathbf{A}$ is skew hermitian (or real skew symmetric), then

$$\mathbf{U} = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$$

is unitary (orthogonal) by first showing that $(\mathbf{I}+\mathbf{A})^{-1}$ exists for skew-hermitian matrices, and $(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$ (recall Exercise 3.7.6). **Note:** There is a more direct approach, but it requires the diagonalization theorem for normal matrices—see Exercise 7.5.5.

5.6.7. Suppose that $\mathbf{R}$ and $\mathbf{S}$ are elementary reflectors.
(a) Is $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$ an elementary reflector?
(b) Is $\begin{pmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}$ an elementary reflector?

5.6.8. (a) Explain why the standard inner product is invariant under a unitary transformation. That is, if $\mathbf{U}$ is any unitary matrix, and if $\mathbf{u} = \mathbf{U}\mathbf{x}$ and $\mathbf{v} = \mathbf{U}\mathbf{y}$, then

$$\mathbf{u}^*\mathbf{v} = \mathbf{x}^*\mathbf{y}.$$

(b) Given any two vectors $\mathbf{x}, \mathbf{y} \in \Re^n$, explain why the angle between them is invariant under an orthogonal transformation. That is, if $\mathbf{u} = \mathbf{P}\mathbf{x}$ and $\mathbf{v} = \mathbf{P}\mathbf{y}$, where $\mathbf{P}$ is an orthogonal matrix, then

$$\cos\theta_{\mathbf{u},\mathbf{v}} = \cos\theta_{\mathbf{x},\mathbf{y}}.$$

5.6.9. Let $\mathbf{U}_{m\times r}$ be a matrix with orthonormal columns, and let $\mathbf{V}_{k\times n}$ be a matrix with orthonormal rows. For an arbitrary $\mathbf{A} \in \mathcal{C}^{r\times k}$, solve the following problems using the matrix 2-norm (p. 281) and the Frobenius matrix norm (p. 279).

(a) Determine the values of $\|\mathbf{U}\|_2$, $\|\mathbf{V}\|_2$, $\|\mathbf{U}\|_F$, and $\|\mathbf{V}\|_F$.

(b) Show that $\|\mathbf{U}\mathbf{A}\mathbf{V}\|_2 = \|\mathbf{A}\|_2$. (**Hint:** Start with $\|\mathbf{U}\mathbf{A}\|_2$.)

(c) Show that $\|\mathbf{U}\mathbf{A}\mathbf{V}\|_F = \|\mathbf{A}\|_F$.

Note: In particular, these properties are valid when $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. Because of parts (b) and (c), the 2-norm and the F-norm are said to be ***unitarily invariant norms.***

5.6.10. Let $\mathbf{u} = \begin{pmatrix} -2 \\ 1 \\ 3 \\ -1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ 4 \\ 0 \\ -1 \end{pmatrix}$.

(a) Determine the orthogonal projection of $\mathbf{u}$ onto $span\{\mathbf{v}\}$.

(b) Determine the orthogonal projection of $\mathbf{v}$ onto $span\{\mathbf{u}\}$.

(c) Determine the orthogonal projection of $\mathbf{u}$ onto $\mathbf{v}^\perp$.

(d) Determine the orthogonal projection of $\mathbf{v}$ onto $\mathbf{u}^\perp$.

5.6.11. Consider elementary orthogonal projectors $\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^*$.

(a) Prove that $\mathbf{Q}$ is singular.

(b) Now prove that if $\mathbf{Q}$ is $n \times n$, then $rank(\mathbf{Q}) = n - 1$.
Hint: Recall Exercise 4.4.10.

5.6.12. For vectors $\mathbf{u}, \mathbf{x} \in \mathcal{C}^n$ such that $\|\mathbf{u}\| = 1$, let $\mathbf{p}$ be the orthogonal projection of $\mathbf{x}$ onto $span\{\mathbf{u}\}$. Explain why $\|\mathbf{p}\| \le \|\mathbf{x}\|$ with equality holding if and only if $\mathbf{x}$ is a scalar multiple of $\mathbf{u}$.

5.6.13. Let $\mathbf{x} = (1/3)\begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix}$.

(a) Determine an elementary reflector $\mathbf{R}$ such that $\mathbf{Rx}$ lies on the x-axis.

(b) Verify by direct computation that your reflector $\mathbf{R}$ is symmetric, orthogonal, and involutory.

(c) Extend $\mathbf{x}$ to an orthonormal basis for $\Re^3$ by using an elementary reflector.

5.6.14. Let $\mathbf{R} = \mathbf{I} - 2\mathbf{uu}^*$, where $\|\mathbf{u}_{n\times 1}\| = 1$. If $\mathbf{x}$ is a *fixed point* for $\mathbf{R}$ in the sense that $\mathbf{Rx} = \mathbf{x}$, and if $n > 1$, prove that $\mathbf{x}$ must be orthogonal to $\mathbf{u}$, and then sketch a picture of this situation in $\Re^3$.

5.6.15. Let $\mathbf{x}, \mathbf{y} \in \Re^{n\times 1}$ be vectors such that $\|\mathbf{x}\| = \|\mathbf{y}\|$ but $\mathbf{x} \neq \mathbf{y}$. Explain how to construct an elementary reflector $\mathbf{R}$ such that $\mathbf{Rx} = \mathbf{y}$.
Hint: The vector $\mathbf{u}$ that defines $\mathbf{R}$ can be determined visually in $\Re^3$ by considering Figure 5.6.2.

5.6.16. Let $\mathbf{x}_{n\times 1}$ be a vector such that $\|\mathbf{x}\| = 1$, and partition $\mathbf{x}$ as

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \tilde{\mathbf{x}} \end{pmatrix}, \qquad \text{where } \tilde{\mathbf{x}} \text{ is } n-1 \times 1.$$

(a) If the entries of $\mathbf{x}$ are real, and if $x_1 \neq 1$, show that

$$\mathbf{P} = \begin{pmatrix} x_1 & \tilde{\mathbf{x}}^T \\ \tilde{\mathbf{x}} & \mathbf{I} - \alpha\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \end{pmatrix}, \quad \text{where} \quad \alpha = \frac{1}{1 - x_1}$$

is an orthogonal matrix.

(b) Suppose that the entries of $\mathbf{x}$ are complex. If $|x_1| \neq 1$, and if μ is the number defined in (5.6.10), show that the matrix

$$\mathbf{U} = \begin{pmatrix} x_1 & \mu^2\tilde{\mathbf{x}}^* \\ \tilde{\mathbf{x}} & \mu(\mathbf{I} - \alpha\tilde{\mathbf{x}}\tilde{\mathbf{x}}^*) \end{pmatrix}, \quad \text{where} \quad \alpha = \frac{1}{1 - |x_1|}$$

is unitary. **Note:** These results provide an easy way to extend a given vector to an orthonormal basis for the entire space $\Re^n$ or $\mathcal{C}^n$.

5.6.17. Perform the following sequence of rotations in $\Re^3$ beginning with

$$\mathbf{v}_0 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

1. Rotate $\mathbf{v}_0$ counterclockwise $45°$ around the x-axis to produce $\mathbf{v}_1$.
2. Rotate $\mathbf{v}_1$ clockwise $90°$ around the y-axis to produce $\mathbf{v}_2$.
3. Rotate $\mathbf{v}_2$ counterclockwise $30°$ around the z-axis to produce $\mathbf{v}_3$.

Determine the coordinates of $\mathbf{v}_3$ as well as an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{v}_0 = \mathbf{v}_3$.

5.6.18. Does it matter in what order rotations in $\Re^3$ are performed? For example, suppose that a vector $\mathbf{v} \in \Re^3$ is first rotated counterclockwise around the x-axis through an angle θ, and then that vector is rotated counterclockwise around the y-axis through an angle ϕ. Is the result the same as first rotating $\mathbf{v}$ counterclockwise around the y-axis through an angle ϕ followed by a rotation counterclockwise around the x-axis through an angle θ?

5.6.19. For each nonzero vector $\mathbf{u} \in \mathcal{C}^n$, prove that $\dim \mathbf{u}^\perp = n - 1$.

5.6.20. A matrix satisfying $\mathbf{A}^2 = \mathbf{I}$ is said to be an ***involution*** or an ***involutory matrix***, and a matrix $\mathbf{P}$ satisfying $\mathbf{P}^2 = \mathbf{P}$ is called a ***projector*** or is said to be an ***idempotent matrix***—properties of such matrices are developed on p. 386. Show that there is a one-to-one correspondence between the set of involutions and the set of projectors in $\mathcal{C}^{n\times n}$. **Hint:** Consider the relationship between the projectors in (5.6.6) and the reflectors (which are involutions) in (5.6.7) on p. 324.

5.6.21. When using a computer to generate and display a three-dimensional convex polytope such as the one in Example 5.6.4, it is desirable to not draw those faces that should be hidden from the perspective of a viewer positioned as shown in Figure 5.6.6. The operation of ***cross product*** in $\Re^3$ (usually introduced in elementary calculus courses) can be used to decide which faces are visible and which are not. Recall that if

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \text{then} \quad \mathbf{u} \times \mathbf{v} = \begin{pmatrix} u_2v_3 - u_3v_2 \\ u_3v_1 - u_1v_3 \\ u_1v_2 - u_2v_1 \end{pmatrix},$$

and $\mathbf{u} \times \mathbf{v}$ is a vector orthogonal to both $\mathbf{u}$ and $\mathbf{v}$. The direction of $\mathbf{u} \times \mathbf{v}$ is determined from the so-called right-hand rule as illustrated in Figure 5.6.8.

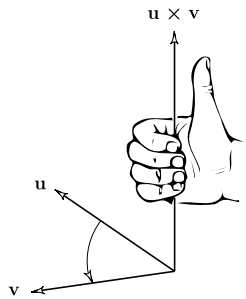


FIGURE 5.6.8

Assume the origin is interior to the polytope, and consider a particular face and three vertices $\mathbf{p}_0$, $\mathbf{p}_1$, and $\mathbf{p}_2$ on the face that are positioned as shown in Figure 5.6.9. The vector $\mathbf{n} = (\mathbf{p}_1 - \mathbf{p}_0) \times (\mathbf{p}_2 - \mathbf{p}_1)$ is orthogonal to the face, and it points in the outward direction.

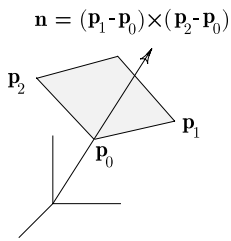


FIGURE 5.6.9

Explain why the outside of the face is visible from the perspective indicated in Figure 5.6.6 if and only if the first component of the outward normal vector $\mathbf{n}$ is positive. In other words, the face is drawn if and only if $n_1 > 0$.

5.7 ORTHOGONAL REDUCTION

We know that a matrix $\mathbf{A}$ can be reduced to row echelon form by elementary row operations. This is Gaussian elimination, and, as explained on p. 143, the basic "Gaussian transformation" is an elementary lower triangular matrix $\mathbf{T}_k$ whose action annihilates all entries below the kth pivot at the kth elimination step. But Gaussian elimination is not the only way to reduce a matrix. Elementary reflectors $\mathbf{R}_k$ can be used in place of elementary lower triangular matrices $\mathbf{T}_k$ to annihilate all entries below the kth pivot at the kth elimination step, or a sequence of plane rotation matrices can accomplish the same purpose.

When reflectors are used, the process is usually called ***Householder reduction***, and it proceeds as follows. For $\mathbf{A}_{m\times n} = [\mathbf{A}_{*1} \,|\, \mathbf{A}_{*2} \,|\, \cdots \,|\, \mathbf{A}_{*n}]$, use $\mathbf{x} = \mathbf{A}_{*1}$ in (5.6.10) to construct the elementary reflector

$$\mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}, \quad \text{where} \quad \mathbf{u} = \mathbf{A}_{*1} \pm \mu \,\|\mathbf{A}_{*1}\|\, \mathbf{e}_1, \tag{5.7.1}$$

so that

$$\mathbf{R}_1\mathbf{A}_{*1} = \mp\mu \,\|\mathbf{A}_{*1}\|\, \mathbf{e}_1 = \begin{pmatrix} t_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.7.2}$$

Applying $\mathbf{R}_1$ to $\mathbf{A}$ yields

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1} \,|\, \mathbf{R}_1\mathbf{A}_{*2} \,|\, \cdots \,|\, \mathbf{R}_1\mathbf{A}_{*n}] = \left(\begin{array}{c|ccc} t_{11} & t_{12} & \cdots & t_{1n} \\ \hline 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{array}\right) = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix},$$

where $\mathbf{A}_2$ is $m-1 \times n-1$. Thus all entries below the $(1,1)$-position are annihilated. Now apply the same procedure to $\mathbf{A}_2$ to construct an elementary reflector $\hat{\mathbf{R}}_2$ that annihilates all entries below the $(1,1)$-position in $\mathbf{A}_2$. If we set $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$, then $\mathbf{R}_2\mathbf{R}_1$ is an orthogonal matrix (Exercise 5.6.5) such that

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} t_{11} & \mathbf{t}_1^T \\ \mathbf{0} & \hat{\mathbf{R}}_2\mathbf{A}_2 \end{pmatrix} = \left(\begin{array}{c|c|ccc} t_{11} & t_{12} & t_{13} & \cdots & t_{1n} \\ \hline 0 & t_{22} & t_{23} & \cdots & t_{2n} \\ \hline 0 & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & * & \cdots & * \end{array}\right).$$

The result after $k-1$ steps is $\mathbf{R}_{k-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} \mathbf{T}_{k-1} & \tilde{\mathbf{T}}_{k-1} \\ \mathbf{0} & \mathbf{A}_k \end{pmatrix}$. At step k an elementary reflector $\hat{\mathbf{R}}_k$ is constructed in a manner similar to (5.7.1)

to annihilate all entries below the $(1,1)$-position in $\mathbf{A}_k$, and $\mathbf{R}_k$ is defined as $\mathbf{R}_k = \begin{pmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_k \end{pmatrix}$, which is another elementary reflector (Exercise 5.6.7). Eventually, all of the rows or all of the columns will be exhausted, so the final result is one of the two following ***upper-trapezoidal forms***:

$$\mathbf{R}_n \cdots \mathbf{R}_2 \mathbf{R}_1 \mathbf{A}_{m\times n} = \left(\begin{array}{cccc} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & * \\ \hline 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}*\\0\\\vdots\\0\end{array}}\right\} n\times n \\ \\ \end{array} \qquad \text{when} \quad m > n,$$

$$\mathbf{R}_{m-1} \cdots \mathbf{R}_2 \mathbf{R}_1 \mathbf{A}_{m\times n} = \left(\begin{array}{cccc|ccc} * & * & \cdots & * & * & \cdots & * \\ 0 & * & \cdots & * & * & \cdots & * \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & * & * & \cdots & * \end{array}\right) \qquad \text{when} \quad m < n.$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{m\times m}$$

If $m = n$, then the final form is an upper-triangular matrix.

A product of elementary reflectors is not necessarily another elementary reflector, but a product of unitary (orthogonal) matrices is again unitary (orthogonal) (Exercise 5.6.5). The elementary reflectors $\mathbf{R}_i$ described above are unitary (orthogonal in the real case) matrices, so every product $\mathbf{R}_k \mathbf{R}_{k-1} \cdots \mathbf{R}_2 \mathbf{R}_1$ is a unitary matrix, and thus we arrive at the following important conclusion.

Orthogonal Reduction

- For every $\mathbf{A} \in \mathcal{C}^{m\times n}$, there exists a unitary matrix $\mathbf{P}$ such that

$$\mathbf{PA} = \mathbf{T} \tag{5.7.3}$$

  has an upper-trapezoidal form. When $\mathbf{P}$ is constructed as a product of elementary reflectors as described above, the process is called ***Householder reduction.***

- If $\mathbf{A}$ is square, then $\mathbf{T}$ is upper triangular, and if $\mathbf{A}$ is real, then the $\mathbf{P}$ can be taken to be an orthogonal matrix.

Example 5.7.1

Problem: Use Householder reduction to find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{T}$ is upper triangular with positive diagonal entries, where

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}.$$

Solution: To annihilate the entries below the $(1,1)$-position and to guarantee that t_{11} is positive, equations (5.7.1) and (5.7.2) dictate that we set

$$\mathbf{u}_1 = \mathbf{A}_{*1} - \|\mathbf{A}_{*1}\|\,\mathbf{e}_1 = \mathbf{A}_{*1} - 5\mathbf{e}_1 = \begin{pmatrix} -5 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{u}_1\mathbf{u}_1^T}{\mathbf{u}_1^T\mathbf{u}_1}.$$

To compute a reflector-by-matrix product $\mathbf{RA} = [\mathbf{RA}_{*1} \,|\, \mathbf{RA}_{*2} \,|\, \cdots \,|\, \mathbf{RA}_{*n}]$, it's wasted effort to actually determine the entries in $\mathbf{R} = \mathbf{I} - 2\mathbf{uu}^T/\mathbf{u}^T\mathbf{u}$. Simply compute $\mathbf{u}^T\mathbf{A}_{*j}$ and then

$$\mathbf{RA}_{*j} = \mathbf{A}_{*j} - 2\left(\frac{\mathbf{u}^T\mathbf{A}_{*j}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u} \quad \text{for each } j = 1, 2, \ldots, n. \tag{5.7.4}$$

By using this observation we obtain

$$\mathbf{R}_1\mathbf{A} = [\mathbf{R}_1\mathbf{A}_{*1} \,|\, \mathbf{R}_1\mathbf{A}_{*2} \,|\, \mathbf{R}_1\mathbf{A}_{*3}] = \left(\begin{array}{c|cc} 5 & 25 & -4 \\ \hline 0 & 0 & -10 \\ 0 & -25 & -10 \end{array}\right).$$

To annihilate the entry below the $(2,2)$-position, set

$$\mathbf{A}_2 = \begin{pmatrix} 0 & -10 \\ -25 & -10 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = [\mathbf{A}_2]_{*1} - \Big\| [\mathbf{A}_2]_{*1} \Big\| \mathbf{e}_1 = 25\begin{pmatrix} -1 \\ -1 \end{pmatrix}.$$

If $\hat{\mathbf{R}}_2 = \mathbf{I} - 2\mathbf{u}_2\mathbf{u}_2^T/\mathbf{u}_2^T\mathbf{u}_2$ and $\mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$ (neither is explicitly computed), then

$$\hat{\mathbf{R}}_2\mathbf{A}_2 = \begin{pmatrix} 25 & 10 \\ 0 & 10 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_2\mathbf{R}_1\mathbf{A} = \mathbf{T} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

If $\hat{\mathbf{R}}_k = \mathbf{I} - 2\hat{\mathbf{u}}\hat{\mathbf{u}}^T/\hat{\mathbf{u}}^T\hat{\mathbf{u}}$ is an elementary reflector, then so is

$$\mathbf{R}_k = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_k \end{pmatrix} = \mathbf{I} - 2\frac{\mathbf{uu}^T}{\mathbf{u}^T\mathbf{u}} \quad \text{with} \quad \mathbf{u} = \begin{pmatrix} \mathbf{0} \\ \hat{\mathbf{u}} \end{pmatrix},$$

and consequently the product of any sequence of these $\mathbf{R}_k$'s can be formed by using the observation (5.7.4). In this example,

$$\mathbf{P} = \mathbf{R}_2\mathbf{R}_1 = \frac{1}{25}\begin{pmatrix} 0 & 15 & 20 \\ -20 & 12 & -9 \\ -15 & -16 & 12 \end{pmatrix}.$$

You may wish to check that $\mathbf{P}$ really is an orthogonal matrix and $\mathbf{PA} = \mathbf{T}$.

Elementary reflectors are not the only type of orthogonal matrices that can be used to reduce a matrix to an upper-trapezoidal form. Plane rotation matrices are also orthogonal, and, as explained on p. 334, plane rotation matrices can be used to selectively annihilate any component in a given column, so a sequence of plane rotations can be used to annihilate all elements below a particular pivot. This means that a matrix $\mathbf{A} \in \Re^{m\times n}$ can be reduced to an upper-trapezoidal form strictly by using plane rotations—such a process is usually called a ***Givens reduction.***

Example 5.7.2

Problem: Use Givens reduction (i.e., use plane rotations) to reduce the matrix

$$\mathbf{A} = \begin{pmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{pmatrix}$$

to upper-triangular form. Also compute an orthogonal matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{T}$ is upper triangular.

Solution: The plane rotation that uses the (1,1)-entry to annihilate the (2,1)-entry is determined from (5.6.16) to be

$$\mathbf{P}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{12}\mathbf{A} = \begin{pmatrix} 3 & 27 & -4 \\ 0 & 20 & 14 \\ 4 & 11 & -2 \end{pmatrix}.$$

Now use the (1,1)-entry in $\mathbf{P}_{12}\mathbf{A}$ to annihilate the (3,1)-entry in $\mathbf{P}_{12}\mathbf{A}$. The plane rotation that does the job is again obtained from (5.6.16) to be

$$\mathbf{P}_{13} = \frac{1}{5}\begin{pmatrix} 3 & 0 & 4 \\ 0 & 5 & 0 \\ -4 & 0 & 3 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 20 & 14 \\ 0 & -15 & 2 \end{pmatrix}.$$

Finally, using the (2,2)-entry in $\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A}$ to annihilate the (3,2)-entry produces

$$\mathbf{P}_{23} = \frac{1}{5}\begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & -3 \\ 0 & 3 & 4 \end{pmatrix} \quad \text{so that} \quad \mathbf{P}_{23}\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A} = \mathbf{T} = \begin{pmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{pmatrix}.$$

Since plane rotation matrices are orthogonal, and since the product of orthogonal matrices is again orthogonal, it must be the case that

$$\mathbf{P} = \mathbf{P}_{23}\mathbf{P}_{13}\mathbf{P}_{12} = \frac{1}{25}\begin{pmatrix} 0 & 15 & 20 \\ -20 & 12 & -9 \\ -15 & -16 & 12 \end{pmatrix}$$

is an orthogonal matrix such that $\mathbf{PA} = \mathbf{T}$.

Householder and Givens reductions are closely related to the results produced by applying the Gram–Schmidt process (p. 307) to the columns of $\mathbf{A}$. When $\mathbf{A}$ is nonsingular, Householder, Givens, and Gram–Schmidt each produce an orthogonal matrix $\mathbf{Q}$ and an upper-triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$ ($\mathbf{Q} = \mathbf{P}^T$ in the case of orthogonal reduction). The upper-triangular matrix $\mathbf{R}$ produced by the Gram–Schmidt algorithm has positive diagonal entries, and, as illustrated in Examples 5.7.1 and 5.7.2, we can also force this to be true using the Householder or Givens reduction. This feature makes $\mathbf{Q}$ and $\mathbf{R}$ unique.

QR Factorization

For each nonsingular $\mathbf{A} \in \Re^{n\times n}$, there is a unique orthogonal matrix $\mathbf{Q}$ and a unique upper-triangular matrix $\mathbf{R}$ with positive diagonal entries such that

$$\mathbf{A} = \mathbf{QR}.$$

This "square" QR factorization is a special case of the more general "rectangular" QR factorization discussed on p. 311.

Proof. Only uniqueness needs to be proven. If there are two QR factorizations

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2,$$

let $\mathbf{U} = \mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{R}_2\mathbf{R}_1^{-1}$. The matrix $\mathbf{R}_2\mathbf{R}_1^{-1}$ is upper triangular with positive diagonal entries (Exercises 3.5.8 and 3.7.4) while $\mathbf{Q}_2^T\mathbf{Q}_1$ is an orthogonal matrix (Exercise 5.6.5), and therefore $\mathbf{U}$ is an upper-triangular matrix whose columns are an orthonormal set and whose diagonal entries are positive. Considering the first column of $\mathbf{U}$ we see that

$$\left\| \begin{pmatrix} u_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\| = 1 \implies u_{11} = \pm 1 \quad \text{and} \quad u_{11} > 0 \implies u_{11} = 1,$$

so that $\mathbf{U}_{*1} = \mathbf{e}_1$. A similar argument together with the fact that the columns of $\mathbf{U}$ are mutually orthogonal produces

$$\mathbf{U}_{*1}^T\mathbf{U}_{*2} = 0 \implies u_{12} = 0 \implies u_{22} = 1 \implies \mathbf{U}_{*2} = \mathbf{e}_2.$$

Proceeding inductively establishes that $\mathbf{U}_{*k} = \mathbf{e}_k$ for each k (i.e., $\mathbf{U} = \mathbf{I}$), and therefore $\mathbf{Q}_1 = \mathbf{Q}_2$ and $\mathbf{R}_1 = \mathbf{R}_2$. ∎

Example 5.7.3

Orthogonal Reduction and Least Squares. Orthogonal reduction can be used to solve the least squares problem associated with an inconsistent system $\mathbf{Ax} = \mathbf{b}$ in which $\mathbf{A} \in \Re^{m\times n}$ and $m \geq n$ (the most common case). If $\boldsymbol{\varepsilon}$ denotes the difference $\boldsymbol{\varepsilon} = \mathbf{Ax} - \mathbf{b}$, then, as described on p. 226, the general least squares problem is to find a vector $\mathbf{x}$ that minimizes the quantity

$$\sum_{i=1}^{m} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \|\boldsymbol{\varepsilon}\|^2,$$

where $\|\star\|$ is the standard euclidean vector norm. Suppose that $\mathbf{A}$ is reduced to an upper-trapezoidal matrix $\mathbf{T}$ by an orthogonal matrix $\mathbf{P}$, and write

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R}_{n\times n} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{Pb} = \begin{pmatrix} \mathbf{c}_{n\times 1} \\ \mathbf{d} \end{pmatrix}$$

in which $\mathbf{R}$ is an upper-triangular matrix. An orthogonal matrix is an isometry—recall (5.6.1)—so that

$$\|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}\boldsymbol{\varepsilon}\|^2 = \|\mathbf{P}(\mathbf{Ax}-\mathbf{b})\|^2 = \left\| \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \mathbf{x} - \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \mathbf{Rx}-\mathbf{c} \\ \mathbf{d} \end{pmatrix} \right\|^2$$
$$= \|\mathbf{Rx}-\mathbf{c}\|^2 + \|\mathbf{d}\|^2.$$

Consequently, $\|\boldsymbol{\varepsilon}\|^2$ is minimized when $\mathbf{x}$ is a vector such that $\|\mathbf{Rx}-\mathbf{c}\|^2$ is minimal or, in other words, $\mathbf{x}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\mathbf{x}$ is a least squares solution for $\mathbf{Rx} = \mathbf{c}$.

Full-Rank Case. In a majority of applications the coefficient matrix $\mathbf{A}$ has linearly independent columns so $rank\,(\mathbf{A}_{m\times n}) = n$. Because multiplication by a nonsingular matrix $\mathbf{P}$ does not change the rank,

$$n = rank\,(\mathbf{A}) = rank\,(\mathbf{PA}) = rank\,(\mathbf{T}) = rank\,(\mathbf{R}_{n\times n}).$$

Thus $\mathbf{R}$ is nonsingular, and we have established the following fact.

- If $\mathbf{A}$ has linearly independent columns, then the (unique) least squares solution for $\mathbf{Ax} = \mathbf{b}$ is obtained by solving the nonsingular triangular system $\mathbf{Rx} = \mathbf{c}$ for $\mathbf{x}$.

As pointed out in Example 4.5.1, computing the matrix product $\mathbf{A}^T\mathbf{A}$ is to be avoided when floating-point computation is used because of the possible loss of significant information. Notice that the method based on orthogonal reduction sidesteps this potential problem because the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ are avoided and the product $\mathbf{A}^T\mathbf{A}$ is never explicitly computed. Householder reduction (or Givens reduction for sparse problems) is a numerically stable algorithm (see the discussion following this example) for solving the full-rank least squares problem, and, if the computations are properly ordered, it is an attractive alternative to the method of Example 5.5.3 that is based on the modified Gram–Schmidt procedure.

We now have four different ways to reduce a matrix to an upper-triangular (or trapezoidal) form. (1) Gaussian elimination; (2) Gram–Schmidt procedure; (3) Householder reduction; and (4) Givens reduction. It's natural to try to compare them and to sort out the advantages and disadvantages of each.

First consider numerical stability. This is a complicated issue, but you can nevertheless gain an intuitive feel for the situation by considering the effect of applying a sequence of "elementary reduction" matrices to a small perturbation of $\mathbf{A}$. Let $\mathbf{E}$ be a matrix such that $\|\mathbf{E}\|_F$ is small relative to $\|\mathbf{A}\|_F$ (the Frobenius norm was introduced on p. 279), and consider

$$\mathbf{P}_k \cdots \mathbf{P}_2\mathbf{P}_1(\mathbf{A}+\mathbf{E}) = (\mathbf{P}_k \cdots \mathbf{P}_2\mathbf{P}_1\mathbf{A}) + (\mathbf{P}_k \cdots \mathbf{P}_2\mathbf{P}_1\mathbf{E}) = \mathbf{PA} + \mathbf{PE}.$$

If each $\mathbf{P}_i$ is an orthogonal matrix, then the product $\mathbf{P} = \mathbf{P}_k \cdots \mathbf{P}_2\mathbf{P}_1$ is also an orthogonal matrix (Exercise 5.6.5), and consequently $\|\mathbf{PE}\|_F = \|\mathbf{E}\|_F$ (Exercise 5.6.9). In other words, a sequence of orthogonal transformations cannot magnify the magnitude of $\mathbf{E}$, and you might think of $\mathbf{E}$ as representing the effects of roundoff error. This suggests that Householder and Givens reductions should be numerically stable algorithms. On the other hand, if the $\mathbf{P}_i$'s are elementary matrices of Type I, II, or III, then the product $\mathbf{P} = \mathbf{P}_k \cdots \mathbf{P}_2\mathbf{P}_1$ can be any nonsingular matrix—recall (3.9.3). Nonsingular matrices are not generally norm preserving (i.e., it is possible that $\|\mathbf{PE}\|_F > \|\mathbf{E}\|_F$), so the possibility of $\mathbf{E}$ being magnified is generally present in elimination methods, and this suggests the possibility of numerical instability.

Strictly speaking, an algorithm is considered to be ***numerically stable*** if, under floating-point arithmetic, it always returns an answer that is the exact solution of a nearby problem. To give an intuitive argument that the Householder or Givens reduction is a stable algorithm for producing the QR factorization of $\mathbf{A}_{n\times n}$, suppose that $\mathbf{Q}$ and $\mathbf{R}$ are the exact QR factors, and suppose that floating-point arithmetic produces an orthogonal matrix $\mathbf{Q}+\mathbf{E}$ and an upper-triangular matrix $\mathbf{R}+\mathbf{F}$ that are the exact QR factors of a different matrix

$$\tilde{\mathbf{A}} = (\mathbf{Q}+\mathbf{E})(\mathbf{R}+\mathbf{F}) = \mathbf{QR} + \mathbf{QF} + \mathbf{ER} + \mathbf{EF} = \mathbf{A} + \mathbf{QF} + \mathbf{ER} + \mathbf{EF}.$$

If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, then the entries in $\mathbf{EF}$ are negligible, and

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{QF} + \mathbf{ER}.$$

But since $\mathbf{Q}$ is orthogonal, $\|\mathbf{QF}\|_F = \|\mathbf{F}\|_F$ and $\|\mathbf{A}\|_F = \|\mathbf{QR}\|_F = \|\mathbf{R}\|_F$, and this means that neither $\mathbf{QF}$ nor $\mathbf{ER}$ can contain entries that are large relative to those in $\mathbf{A}$. Hence $\tilde{\mathbf{A}} \approx \mathbf{A}$, and this is what is required to conclude that the algorithm is stable.

Gaussian elimination is not a stable algorithm because, as alluded to in §1.5, problems arise due to the growth of the magnitude of the numbers that can occur

during the process. To see this from a heuristic point of view, consider the LU factorization of $\mathbf{A} = \mathbf{LU}$, and suppose that floating-point Gaussian elimination with no pivoting returns matrices $\mathbf{L} + \mathbf{E}$ and $\mathbf{U} + \mathbf{F}$ that are the exact LU factors of a somewhat different matrix

$$\tilde{\mathbf{A}} = (\mathbf{L} + \mathbf{E})(\mathbf{U} + \mathbf{F}) = \mathbf{LU} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF} = \mathbf{A} + \mathbf{LF} + \mathbf{EU} + \mathbf{EF}.$$

If $\mathbf{E}$ and $\mathbf{F}$ account for the roundoff errors, and if their entries are small relative to those in $\mathbf{A}$, then the entries in $\mathbf{EF}$ are negligible, and

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{LF} + \mathbf{EU} \qquad \text{(using no pivoting)}.$$

However, if $\mathbf{L}$ or $\mathbf{U}$ contains entries that are large relative to those in $\mathbf{A}$ (and this is certainly possible), then $\mathbf{LF}$ or $\mathbf{EU}$ can contain entries that are significant. In other words, Gaussian elimination with no pivoting can return the LU factorization of a matrix $\tilde{\mathbf{A}}$ that is not very close to the original matrix $\mathbf{A}$, and this is what it means to say that an algorithm is unstable. We saw on p. 26 that if partial pivoting is employed, then no multiplier can exceed 1 in magnitude, and hence no entry of $\mathbf{L}$ can be greater than 1 in magnitude (recall that the subdiagonal entries of $\mathbf{L}$ are in fact the multipliers). Consequently, $\mathbf{L}$ cannot greatly magnify the entries of $\mathbf{F}$, so, if the rows of $\mathbf{A}$ have been reordered according to the partial pivoting strategy, then

$$\tilde{\mathbf{A}} \approx \mathbf{A} + \mathbf{EU} \qquad \text{(using partial pivoting)}.$$

Numerical stability requires that $\tilde{\mathbf{A}} \approx \mathbf{A}$, so the issue boils down to the degree to which $\mathbf{U}$ magnifies the entries in $\mathbf{E}$—i.e., the issue rests on the magnitude of the entries in $\mathbf{U}$. Unfortunately, partial pivoting may not be enough to control the growth of all entries in $\mathbf{U}$. For example, when Gaussian elimination with partial pivoting is applied to

$$\mathbf{W}_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -1 & 1 & \ddots & 0 & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \ddots & 1 & 0 & 1 \\ -1 & -1 & -1 & \cdots & -1 & 1 & 1 \\ -1 & -1 & -1 & \cdots & -1 & -1 & 1 \end{pmatrix},$$

the largest entry in $\mathbf{U}$ is $u_{nn} = 2^{n-1}$. However, if complete pivoting is used on $\mathbf{W}_n$, then no entry in the process exceeds 2 in magnitude (Exercises 1.5.7 and 1.5.8). In general, it has been proven that if complete pivoting is used on a well-scaled matrix $\mathbf{A}_{n\times n}$ for which $\max|a_{ij}| = 1$, then no entry of $\mathbf{U}$ can exceed

$\gamma = n^{1/2}\left(2^1 3^{1/2} 4^{1/3} \cdots n^{1/n-1}\right)^{1/2}$ in magnitude. Since γ is a slow growing function of n, the entries in $\mathbf{U}$ won't greatly magnify the entries of $\mathbf{E}$, so

$$\tilde{\mathbf{A}} \approx \mathbf{A} \qquad \text{(using complete pivoting).}$$

In other words, Gaussian elimination with complete pivoting is stable, but Gaussian elimination with partial pivoting is not. Fortunately, in practical work it is rare to encounter problems such as the matrix $\mathbf{W}_n$ in which partial pivoting fails to control the growth in the $\mathbf{U}$ factor, so scaled partial pivoting is generally considered to be a "practically stable" algorithm.

Algorithms based on the Gram–Schmidt procedure are more complicated. First, the Gram–Schmidt algorithms differ from Householder and Givens reductions in that the Gram–Schmidt procedures are not a sequential application of elementary orthogonal transformations. Second, as an algorithm to produce the QR factorization even the modified Gram–Schmidt technique can return a Q factor that is far from being orthogonal, and the intuitive stability argument used earlier is not valid. As an algorithm to return the QR factorization of $\mathbf{A}$, the modified Gram–Schmidt procedure has been proven to be unstable, but as an algorithm used to solve the least squares problem (see Example 5.5.3), it is stable—i.e., stability of modified Gram–Schmidt is problem dependent.

Summary of Numerical Stability

- Gaussian elimination with scaled partial pivoting is theoretically unstable, but it is "practically stable"—i.e., stable for most practical problems.
- Complete pivoting makes Gaussian elimination unconditionally stable.
- For the QR factorization, the Gram–Schmidt procedure (classical or modified) is not stable. However, the modified Gram–Schmidt procedure is a stable algorithm for solving the least squares problem.
- Householder and Givens reductions are unconditionally stable algorithms for computing the QR factorization.

For the algorithms under consideration, the number of multiplicative operations is about the same as the number of additive operations, so computational effort is gauged by counting only multiplicative operations. For the sake of comparison, lower-order terms are not significant, and when they are neglected the following approximations are obtained.

Summary of Computational Effort

The approximate number of multiplications/divisions required to reduce an $n \times n$ matrix to an upper-triangular form is as follows.

- Gaussian elimination (scaled partial pivoting) $\approx n^3/3$.
- Gram–Schmidt procedure (classical and modified) $\approx n^3$.
- Householder reduction $\approx 2n^3/3$.
- Givens reduction $\approx 4n^3/3$.

It's not surprising that the unconditionally stable methods tend to be more costly—there is no free lunch. No one triangularization technique can be considered optimal, and each has found a place in practical work. For example, in solving unstructured linear systems, the probability of Gaussian elimination with scaled partial pivoting failing is not high enough to justify the higher cost of using the safer Householder or Givens reduction, or even complete pivoting. Although much the same is true for the full-rank least squares problem, Householder reduction or modified Gram–Schmidt is frequently used as a safeguard against sensitivities that often accompany least squares problems. For the purpose of computing an orthonormal basis for $R(\mathbf{A})$ in which $\mathbf{A}$ is unstructured and dense (not many zeros), Householder reduction is preferred—the Gram–Schmidt procedures are unstable for this purpose and Givens reduction is too costly. Givens reduction is useful when the matrix being reduced is highly structured or sparse (many zeros).

Example 5.7.4

Reduction to Hessenberg Form. For reasons alluded to in §4.8 and §4.9, it is often desirable to triangularize a square matrix $\mathbf{A}$ by means of a similarity transformation—i.e., find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{T}$ is upper triangular. But this is a computationally difficult task, so we will try to do the next best thing, which is to find a similarity transformation that will reduce $\mathbf{A}$ to a matrix in which all entries below the first subdiagonal are zero. Such a matrix is said to be in ***upper-Hessenberg form***—illustrated below is a 5×5 Hessenberg form.

$$\mathbf{H} = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

Problem: Reduce $\mathbf{A} \in \Re^{n\times n}$ to upper-Hessenberg form by means of an orthogonal similarity transformation—i.e., construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}$ is upper Hessenberg.

Solution: At each step, use Householder reduction on entries *below* the main diagonal. Begin by letting $\hat{\mathbf{A}}_{*1}$ denote the entries of the first column that are below the (1,1)-position—this is illustrated below for $n = 5$:

$$\mathbf{A} = \left(\begin{array}{c|cccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{array}\right) = \left(\begin{array}{c|c} a_{11} & \hat{\mathbf{A}}_{1*} \\ \hline \hat{\mathbf{A}}_{*1} & \mathbf{A}_1 \end{array}\right).$$

If $\hat{\mathbf{R}}_1$ is an elementary reflector determined according to (5.7.1) for which $\hat{\mathbf{R}}_1\hat{\mathbf{A}}_{*1} = \begin{pmatrix} * \\ 0 \\ 0 \\ 0 \end{pmatrix}$, then $\mathbf{R}_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix}$ is an orthogonal matrix such that

$$\mathbf{R}_1\mathbf{A}\mathbf{R}_1 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix} \begin{pmatrix} a_{11} & \hat{\mathbf{A}}_{1*} \\ \hat{\mathbf{A}}_{*1} & \mathbf{A}_1 \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_1 \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & \hat{\mathbf{A}}_{1*}\hat{\mathbf{R}}_1 \\ \hat{\mathbf{R}}_1\hat{\mathbf{A}}_{*1} & \hat{\mathbf{R}}_1\mathbf{A}_1\hat{\mathbf{R}}_1 \end{pmatrix} = \left(\begin{array}{c|cccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{array}\right).$$

At the second step, repeat the process on $\mathbf{A}_2 = \hat{\mathbf{R}}_1\mathbf{A}_1\hat{\mathbf{R}}_1$ to obtain an orthogonal matrix $\hat{\mathbf{R}}_2$ such that $\hat{\mathbf{R}}_2\mathbf{A}_2\hat{\mathbf{R}}_2 = \left(\begin{array}{c|ccc} * & * & * & * \\ \hline * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{array}\right)$. Matrix $\mathbf{R}_2 = \begin{pmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix}$ is an orthogonal matrix such that

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A}\mathbf{R}_1\mathbf{R}_2 = \left(\begin{array}{c|c|ccc} * & * & * & * & * \\ \hline * & * & * & * & * \\ \hline 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{array}\right).$$

After $n-2$ of these steps, the product $\mathbf{P} = \mathbf{R}_1\mathbf{R}_2\cdots\mathbf{R}_{n-2}$ is an orthogonal matrix such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}$ is in upper-Hessenberg form.

Note: If $\mathbf{A}$ is a symmetric matrix, then $\mathbf{H}^T = (\mathbf{P}^T\mathbf{A}\mathbf{P})^T = \mathbf{P}^T\mathbf{A}^T\mathbf{P} = \mathbf{H}$, so $\mathbf{H}$ is symmetric. But as illustrated below for $n = 5$, a symmetric Hessenberg form is a tridiagonal matrix,

$$\mathbf{H} = \mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & * & * & * & 0 \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix},$$

so the following useful corollary is obtained.

- Every real-symmetric matrix is orthogonally similar to a tridiagonal matrix, and Householder reduction can be used to compute this tridiagonal matrix. However, the Lanczos technique discussed on p. 651 can be much more efficient.

Example 5.7.5

Problem: Compute the QR factors of a nonsingular upper-Hessenberg matrix $\mathbf{H} \in \Re^{n\times n}$.

Solution: Due to its smaller multiplication count, Householder reduction is generally preferred over Givens reduction. The exception is for matrices that have a zero pattern that can be exploited by the Givens method but not by the Householder method. A Hessenberg matrix $\mathbf{H}$ is such an example. The first step of Householder reduction completely destroys most of the zeros in $\mathbf{H}$, but applying plane rotations does not. This is illustrated below for a 5×5 Hessenberg form—remember that the action of $\mathbf{P}_{k,k+1}$ affects only the kth and $(k+1)^{st}$ rows.

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{12}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{23}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}$$

$$\xrightarrow{\mathbf{P}_{34}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix} \xrightarrow{\mathbf{P}_{45}} \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * \end{pmatrix}.$$

In general, $\mathbf{P}_{n-1,n}\cdots\mathbf{P}_{23}\mathbf{P}_{12}\mathbf{H} = \mathbf{R}$ is upper triangular in which all diagonal entries, except possibly the last, are positive—the last diagonal can be made positive by the technique illustrated in Example 5.7.2. Thus we obtain an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}\mathbf{H} = \mathbf{R}$, or $\mathbf{H} = \mathbf{Q}\mathbf{R}$ in which $\mathbf{Q} = \mathbf{P}^T$.

Example 5.7.6

Jacobi Reduction.[49] Given a real-symmetric matrix $\mathbf{A}$, the result of Example 5.7.4 shows that Householder reduction can be used to construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{T}$ is tridiagonal. Can we do better?—i.e., can we construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D}$ is a diagonal matrix? Indeed we can, and much of the material in Chapter 7 concerning eigenvalues and eigenvectors is devoted to this problem. But in the present context, this fact can be constructively established by means of Jacobi's diagonalization algorithm.

Jacobi's Idea. If $\mathbf{A} \in \Re^{n\times n}$ is symmetric, then a plane rotation matrix can be applied to reduce the magnitude of the off-diagonal entries. In particular, suppose that $a_{ij} \neq 0$ is the off-diagonal entry of maximal magnitude, and let $\mathbf{A}'$ denote the matrix obtained by setting each $a_{kk} = 0$. If $\mathbf{P}_{ij}$ is the plane rotation matrix described on p. 333 in which $c = \cos\theta$ and $s = \sin\theta$, where $\cot 2\theta = (a_{ii} - a_{jj})/2a_{ij}$, and if $\mathbf{B} = \mathbf{P}_{ij}^T\mathbf{A}\mathbf{P}_{ij}$, then

(1) $b_{ij} = b_{ji} = 0$ (i.e., a_{ij} is annihilated),

(2) $\|\mathbf{B}'\|_F^2 = \|\mathbf{A}'\|_F^2 - 2a_{ij}^2$,

(3) $\|\mathbf{B}'\|_F^2 \leq \left(1 - \dfrac{2}{n^2 - n}\right)\|\mathbf{A}'\|_F^2$.

Proof. The entries of $\mathbf{B} = \mathbf{P}_{ij}^T\mathbf{A}\mathbf{P}_{ij}$ that lay on the intersection of the ith and j^{th} rows with the ith and j^{th} columns can be described by

$$\hat{\mathbf{B}} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \mathbf{P}^T\hat{\mathbf{A}}\mathbf{P}.$$

Use the identities $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\cos\theta\sin\theta$ to verify $b_{ij} = b_{ji} = 0$, and recall that $\|\hat{\mathbf{B}}\|_F = \|\mathbf{P}^T\hat{\mathbf{A}}\mathbf{P}\|_F = \|\hat{\mathbf{A}}\|_F$ (recall Exercise

[49] Karl Gustav Jacob Jacobi (1804–1851) first presented this method in 1846, and it was popular for a time. But the twentieth-century development of electronic computers sparked tremendous interest in numerical algorithms for diagonalizing symmetric matrices, and Jacobi's method quickly fell out of favor because it could not compete with newer procedures—at least on the traditional sequential machines. However, the emergence of multiprocessor parallel computers has resurrected interest in Jacobi's method because of the inherent parallelism in the algorithm. Jacobi was born in Potsdam, Germany, educated at the University of Berlin, and employed as a professor at the University of Königsberg. During his prolific career he made contributions that are still important facets of contemporary mathematics. His accomplishments include the development of elliptic functions; a systematic development and presentation of the theory of determinants; contributions to the theory of rotating liquids; and theorems in the areas of differential equations, calculus of variations, and number theory. In contrast to his great contemporary Gauss, who disliked teaching and was anything but inspiring, Jacobi was regarded as a great teacher (the introduction of the student seminar method is credited to him), and he advocated the view that "the sole end of science is the honor of the human mind, and that under this title a question about numbers is worth as much as a question about the system of the world." Jacobi once defended his excessive devotion to work by saying that "Only cabbages have no nerves, no worries. And what do they get out of their perfect wellbeing?" Jacobi suffered a breakdown from overwork in 1843, and he died at the relatively young age of 46.

5.6.9) to produce the conclusion $b_{ii}^2 + b_{jj}^2 = a_{ii}^2 + 2a_{ij}^2 + a_{jj}^2$. Now use the fact that $b_{kk} = a_{kk}$ for all $k \neq i, j$ together with $\|\mathbf{B}\|_F = \|\mathbf{A}\|_F$ to write

$$\begin{aligned}\|\mathbf{B}'\|_F^2 &= \|\mathbf{B}\|_F^2 - \sum_k b_{kk}^2 = \|\mathbf{B}\|_F^2 - \sum_{k\neq i,j} b_{kk}^2 - \left(b_{ii}^2 + b_{jj}^2\right) \\ &= \|\mathbf{A}\|_F^2 - \sum_{k\neq i,j} a_{kk}^2 - \left(a_{ii}^2 + 2a_{ij}^2 + a_{jj}^2\right) = \|\mathbf{A}\|_F^2 - \sum_k a_{kk}^2 - 2a_{ij}^2 \\ &= \|\mathbf{A}'\|_F^2 - 2a_{ij}^2.\end{aligned}$$

Furthermore, since $a_{pq}^2 \leq a_{ij}^2$ for all $p \neq q$,

$$\|\mathbf{A}'\|_F^2 = \sum_{p\neq q} a_{pq}^2 \leq \sum_{p\neq q} a_{ij}^2 = (n^2 - n)a_{ij}^2 \quad \Longrightarrow \quad -a_{ij}^2 \leq -\frac{\|\mathbf{A}'\|_F^2}{n^2 - n},$$

so

$$\|\mathbf{B}'\|_F^2 = \|\mathbf{A}'\|_F^2 - 2a_{ij}^2 \leq \|\mathbf{A}'\|_F^2 - 2\frac{\|\mathbf{A}'\|_F^2}{n^2 - n} = \left(1 - \frac{2}{n^2 - n}\right)\|\mathbf{A}'\|_F^2. \quad \blacksquare$$

Jacobi's Diagonalization Algorithm. Start with $\mathbf{A}_0 = \mathbf{A}$, and produce a sequence of matrices $\mathbf{A}_k = \mathbf{P}_k^T \mathbf{A}_{k-1} \mathbf{P}_k$, where at the kth step $\mathbf{P}_k$ is a plane rotation constructed to annihilate the maximal off-diagonal entry in $\mathbf{A}_{k-1}$. In particular, if a_{ij} is the entry of maximal magnitude in $\mathbf{A}_{k-1}$, then $\mathbf{P}_k$ is the rotator in the (i, j)-plane defined by setting

$$s = \frac{1}{\sqrt{1+\sigma^2}} \quad \text{and} \quad c = \frac{\sigma}{\sqrt{1+\sigma^2}} = \sqrt{1 - s^2}, \quad \text{where} \quad \sigma = \frac{(a_{ii} - a_{jj})}{2a_{ij}}.$$

For $n > 2$ we have

$$\|\mathbf{A}_k'\|_F^2 \leq \left(1 - \frac{2}{n^2 - n}\right)^k \|\mathbf{A}'\|_F^2 \to 0 \quad \text{as} \quad k \to \infty.$$

Therefore, if $\mathbf{P}^{(k)}$ is the orthogonal matrix defined by $\mathbf{P}^{(k)} = \mathbf{P}_1 \mathbf{P}_2 \cdots \mathbf{P}_k$, then

$$\lim_{k\to\infty} \mathbf{P}^{(k)^T} \mathbf{A} \mathbf{P}^{(k)} = \lim_{k\to\infty} \mathbf{A}_k = \mathbf{D}$$

is a diagonal matrix.

Exercises for section 5.7

5.7.1. (a) Using Householder reduction, compute the QR factors of

$$\mathbf{A} = \begin{pmatrix} 1 & 19 & -34 \\ -2 & -5 & 20 \\ 2 & 8 & 37 \end{pmatrix}.$$

(b) Repeat part (a) using Givens reduction.

5.7.2. For $\mathbf{A} \in \Re^{m\times n}$, suppose that $rank\,(\mathbf{A}) = n$, and let $\mathbf{P}$ be an orthogonal matrix such that

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R}_{n\times n} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{R}$ is an upper-triangular matrix. If $\mathbf{P}^T$ is partitioned as

$$\mathbf{P}^T = [\mathbf{X}_{m\times n} \,|\, \mathbf{Y}],$$

explain why the columns of $\mathbf{X}$ constitute an orthonormal basis for $R\,(\mathbf{A})$.

5.7.3. By using Householder reduction, find an orthonormal basis for $R\,(\mathbf{A})$, where

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 4 \\ 2 & -14 & -3 \\ -2 & 14 & 0 \\ 1 & -7 & 15 \end{pmatrix}.$$

5.7.4. Use Householder reduction to compute the least squares solution for $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 4 \\ 2 & -14 & -3 \\ -2 & 14 & 0 \\ 1 & -7 & 15 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 5 \\ -15 \\ 0 \\ 30 \end{pmatrix}.$$

Hint: Make use of the factors you computed in Exercise 5.7.3.

5.7.5. If $\mathbf{A} = \mathbf{QR}$ is the QR factorization for $\mathbf{A}$, explain why $\|\mathbf{A}\|_F = \|\mathbf{R}\|_F$, where $\|\star\|_F$ is the Frobenius matrix norm introduced on p. 279.

5.7.6. Find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{AP} = \mathbf{H}$ is in upper-Hessenberg form, where

$$\mathbf{A} = \begin{pmatrix} -2 & 3 & -4 \\ 3 & -25 & 50 \\ -4 & 50 & 25 \end{pmatrix}.$$

5.7.7. Let $\mathbf{H}$ be an upper-Hessenberg matrix, and suppose that $\mathbf{H} = \mathbf{QR}$, where $\mathbf{R}$ is a nonsingular upper-triangular matrix. Prove that $\mathbf{Q}$ as well as the product $\mathbf{RQ}$ must also be in upper-Hessenberg form.

5.7.8. Approximately how many multiplications are needed to reduce an $n \times n$ nonsingular upper-Hessenberg matrix to upper-triangular form by using plane rotations?

5.8 DISCRETE FOURIER TRANSFORM

For a positive integer n, the complex numbers $\{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$, where

$$\omega = e^{2\pi i/n} = \cos\frac{2\pi}{n} + i\sin\frac{2\pi}{n}$$

are called the ***n*** th ***roots of unity*** because they represent all solutions to $z^n = 1$. Geometrically, they are the vertices of a regular polygon of n sides as depicted in Figure 5.8.1 for $n = 3$ and $n = 6$.

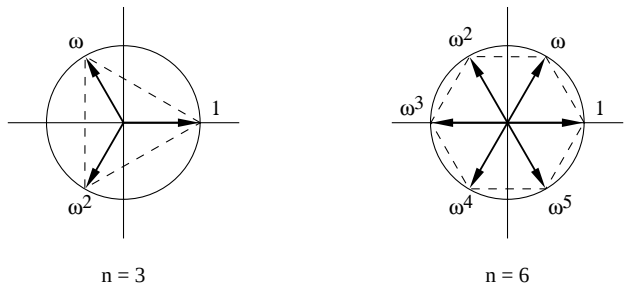


FIGURE 5.8.1

The roots of unity are cyclic in the sense that if $k \geq n$, then $\omega^k = \omega^{k\,(\text{mod } n)}$, where $k\,(\text{mod } n)$ denotes the remainder when k is divided by n—for example, when $n = 6$, $\omega^6 = 1$, $\omega^7 = \omega$, $\omega^8 = \omega^2$, $\omega^9 = \omega^3, \ldots$.

The numbers $\{1, \xi, \xi^2, \ldots, \xi^{n-1}\}$, where

$$\xi = e^{-2\pi i/n} = \cos\frac{2\pi}{n} - i\sin\frac{2\pi}{n} = \overline{\omega}$$

are also the nth roots of unity, but, as depicted in Figure 5.8.2 for $n = 3$ and $n = 6$, they are listed in clockwise order around the unit circle rather than counterclockwise.

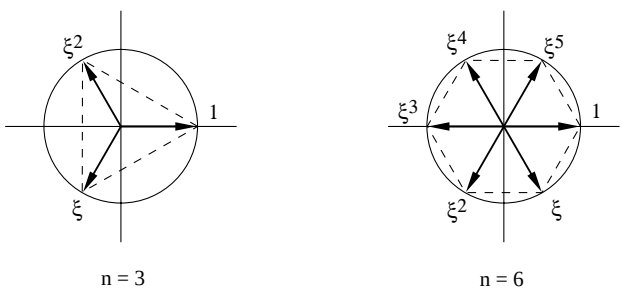


FIGURE 5.8.2

The following identities will be useful in our development. If k is an integer, then $1 = |\xi^k|^2 = \xi^k\overline{\xi^k}$ implies that

$$\xi^{-k} = \overline{\xi^k} = \omega^k. \tag{5.8.1}$$

Furthermore, the fact that

$$\xi^k\left(1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-2)k}+\xi^{(n-1)k}\right)=\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}+1$$

implies $\left(1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}\right)\left(1-\xi^k\right)=0$ and, consequently,

$$1+\xi^k+\xi^{2k}+\cdots+\xi^{(n-1)k}=0 \quad \text{whenever} \quad \xi^k\neq 1. \tag{5.8.2}$$

Fourier Matrix

The $n\times n$ matrix whose (j,k)-entry is $\xi^{jk}=\omega^{-jk}$ for $0\leq j,k\leq n-1$ is called the ***Fourier matrix*** of order n, and it has the form

$$\mathbf{F}_n=\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \xi & \xi^2 & \cdots & \xi^{n-1} \\ 1 & \xi^2 & \xi^4 & \cdots & \xi^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{n-1} & \xi^{n-2} & \cdots & \xi \end{pmatrix}_{n\times n}.$$

Note. Throughout this section entries are indexed from 0 to $n-1$. For example, the upper left-hand entry of $\mathbf{F}_n$ is considered to be in the $(0,0)$ position (rather than the $(1,1)$ position), and the lower right-hand entry is in the $(n-1,n-1)$ position. When the context makes it clear, the subscript n on $\mathbf{F}_n$ is omitted.

The Fourier matrix[50] is a special case of the Vandermonde matrix introduced in Example 4.3.4. Using (5.8.1) and (5.8.2), we see that the inner product of any two columns in $\mathbf{F}_n$, say, the rth and s^{th}, is

$$\mathbf{F}_{*r}^*\mathbf{F}_{*s}=\sum_{j=0}^{n-1}\overline{\xi^{jr}}\xi^{js}=\sum_{j=0}^{n-1}\xi^{-jr}\xi^{js}=\sum_{j=0}^{n-1}\xi^{j(s-r)}=0.$$

In other words, the columns in $\mathbf{F}_n$ are mutually orthogonal. Furthermore, each column in $\mathbf{F}_n$ has norm $\sqrt{n}$ because

$$\|\mathbf{F}_{*k}\|_2^2=\sum_{j=0}^{n-1}|\xi^{jk}|^2=\sum_{j=0}^{n-1}1=n,$$

[50] Some authors define the Fourier matrix using powers of ω rather than powers of ξ, and some include a scalar multiple $1/n$ or $1/\sqrt{n}$. These differences are superficial, and they do not affect the basic properties. Our definition is the discrete counterpart of the integral operator $F(f)=\int_{-\infty}^{\infty}x(t)\mathrm{e}^{-\mathrm{i}2\pi ft}dt$ that is usually taken as the definition of the continuous Fourier transform.

and consequently every column of $\mathbf{F}_n$ can be normalized by multiplying by the same scalar—namely, $1/\sqrt{n}$. This means that $(1/\sqrt{n})\mathbf{F}_n$ is a unitary matrix. Since it is also true that $\mathbf{F}_n^T = \mathbf{F}_n$, we have

$$\left(\frac{1}{\sqrt{n}}\mathbf{F}_n\right)^{-1} = \left(\frac{1}{\sqrt{n}}\mathbf{F}_n\right)^{*} = \frac{1}{\sqrt{n}}\overline{\mathbf{F}}_n,$$

and therefore $\mathbf{F}_n^{-1} = \overline{\mathbf{F}}_n/n$. But (5.8.1) says that $\overline{\xi^k} = \omega^k$, so it must be the case that

$$\mathbf{F}_n^{-1} = \frac{1}{n}\overline{\mathbf{F}}_n = \frac{1}{n}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{n-1} & \omega^{n-2} & \cdots & \omega \end{pmatrix}_{n\times n}.$$

Example 5.8.1

The Fourier matrices of orders 2 and 4 are given by

$$\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix},$$

and their inverses are

$$\mathbf{F}_2^{-1} = \frac{1}{2}\overline{\mathbf{F}}_2 = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{F}_4^{-1} = \frac{1}{4}\overline{\mathbf{F}}_4 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}.$$

Discrete Fourier Transform

Given a vector $\mathbf{x}_{n\times 1}$, the product $\mathbf{F}_n\mathbf{x}$ is called the ***discrete Fourier transform*** of $\mathbf{x}$, and $\mathbf{F}_n^{-1}\mathbf{x}$ is called the ***inverse transform*** of $\mathbf{x}$. The kth entries in $\mathbf{F}_n\mathbf{x}$ and $\mathbf{F}_n^{-1}\mathbf{x}$ are given by

$$[\mathbf{F}_n\mathbf{x}]_k = \sum_{j=0}^{n-1} x_j\xi^{jk} \quad \text{and} \quad [\mathbf{F}_n^{-1}\mathbf{x}]_k = \frac{1}{n}\sum_{j=0}^{n-1} x_j\omega^{jk}. \tag{5.8.3}$$

Example 5.8.2

Problem: Computing the Inverse Transform. Explain why any algorithm or program designed to compute the discrete Fourier transform of a vector $\mathbf{x}$ can also be used to compute the *inverse* transform of $\mathbf{x}$.

Solution: Call such an algorithm FFT (see p. 373 for a specific example). The fact that

$$\mathbf{F}_n^{-1}\mathbf{x} = \frac{\overline{\mathbf{F}}_n\mathbf{x}}{n} = \frac{\overline{\mathbf{F}_n\overline{\mathbf{x}}}}{n}$$

means that FFT will return the inverse transform of $\mathbf{x}$ by executing the following three steps:

(1) $\mathbf{x} \longleftarrow \overline{\mathbf{x}}$ (compute $\overline{\mathbf{x}}$).

(2) $\mathbf{x} \longleftarrow \text{FFT}(\mathbf{x})$ (compute $\mathbf{F}_n\overline{\mathbf{x}}$).

(3) $\mathbf{x} \longleftarrow (1/n)\overline{\mathbf{x}}$ (compute $n^{-1}\overline{\mathbf{F}_n\overline{\mathbf{x}}} = \mathbf{F}_n^{-1}\mathbf{x}$).

For example, computing the inverse transform of $\mathbf{x} = \begin{pmatrix} \mathrm{i} & 0 & -\mathrm{i} & 0 \end{pmatrix}^T$ is accomplished as follows—recall that $\mathbf{F}_4$ was given in Example 5.8.1.

$$\overline{\mathbf{x}} = \begin{pmatrix} -\mathrm{i} \\ 0 \\ \mathrm{i} \\ 0 \end{pmatrix}, \quad \mathbf{F}_4\overline{\mathbf{x}} = \begin{pmatrix} 0 \\ -2\mathrm{i} \\ 0 \\ -2\mathrm{i} \end{pmatrix}, \quad \frac{1}{4}\overline{\mathbf{F}_4\overline{\mathbf{x}}} = \frac{1}{4}\begin{pmatrix} 0 \\ 2\mathrm{i} \\ 0 \\ 2\mathrm{i} \end{pmatrix} = \mathbf{F}_4^{-1}\mathbf{x}.$$

You may wish to check that this answer agrees with the result obtained by directly multiplying $\mathbf{F}_4^{-1}$ times $\mathbf{x}$, where $\mathbf{F}_4^{-1}$ is given in Example 5.8.1.

Example 5.8.3

Signal Processing. Suppose that a microphone is placed under a hovering helicopter, and suppose that Figure 5.8.3 represents the sound signal that is recorded during 1 second of time.

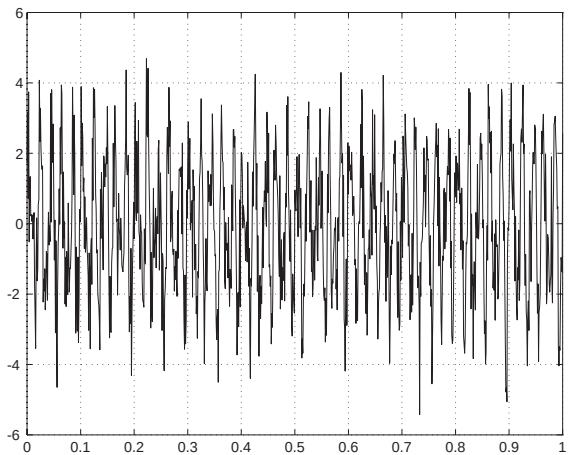


FIGURE 5.8.3

It seems reasonable to expect that the signal should have oscillatory components together with some random noise contamination. That is, we expect the signal to have the form

$$y(\tau) = \left(\sum_k \alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau \right) + \text{Noise}.$$

But due to the noise contamination, the oscillatory nature of the signal is only barely apparent—the characteristic "chop-a chop-a chop-a" is not completely clear. To reveal the oscillatory components, the magic of the Fourier transform is employed. Let $\mathbf{x}$ be the vector obtained by sampling the signal at n equally spaced points between time $\tau = 0$ and $\tau = 1$ ($n = 512$ in our case), and let

$$\mathbf{y} = (2/n)\mathbf{F}_n\mathbf{x} = \mathbf{a} + \mathrm{i}\mathbf{b}, \quad \text{where} \quad \mathbf{a} = (2/n)\mathrm{Re}\,(\mathbf{F}_n\mathbf{x}) \text{ and } \mathbf{b} = (2/n)\mathrm{Im}\,(\mathbf{F}_n\mathbf{x})\,.$$

Using only the first $n/2 = 256$ entries in $\mathbf{a}$ and $\mathrm{i}\mathbf{b}$, we plot the points in

$$\{(0, a_0),\ (1, a_1),\ \ldots,\ (255, a_{255})\} \quad \text{and} \quad \{(0, \mathrm{i}b_0),\ (1, \mathrm{i}b_1),\ \ldots,\ (255, \mathrm{i}b_{255})\}$$

to produce the two graphs shown in Figure 5.8.4.

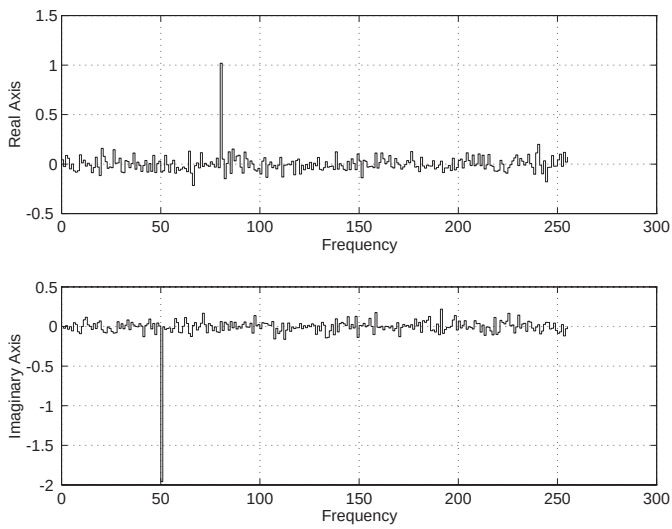


FIGURE 5.8.4

Now there are some obvious characteristics—the plot of $\mathbf{a}$ in the top graph of Figure 5.8.4 has a spike of height approximately 1 at entry 80, and the plot of $\mathrm{i}\mathbf{b}$ in the bottom graph has a spike of height approximately -2 at entry 50. These two spikes indicate that the signal is made up primarily of two oscillatory

components—the spike in the real vector $\mathbf{a}$ indicates that one of the oscillatory components is a cosine of frequency 80 Hz (or period $= 1/80$) whose amplitude is approximately 1, and the spike in the imaginary vector $\mathrm{i}\mathbf{b}$ indicates there is a sine component with frequency 50 Hz and amplitude of about 2. In other words, the Fourier transform indicates that the signal is

$$y(\tau) = \cos 2\pi(80\tau) + 2\sin 2\pi(50\tau) + \text{Noise}.$$

In truth, the data shown in Figure 5.8.3 was artificially generated by contaminating the function $y(\tau) = \cos 2\pi(80\tau) + 2\sin 2\pi(50\tau)$ with some normally distributed zero-mean noise, and therefore the plot of $(2/n)\mathbf{F}_n\mathbf{x}$ shown in Figure 5.8.4 does indeed accurately reflect the true nature of the signal. To understand why $\mathbf{F}_n$ reveals the hidden frequencies, let $\cos 2\pi f\mathbf{t}$ and $\sin 2\pi f\mathbf{t}$ denote the ***discrete cosine*** and ***discrete sine*** vectors

$$\cos 2\pi f\mathbf{t} = \begin{pmatrix} \cos\left(2\pi f\cdot\frac{0}{n}\right) \\ \cos\left(2\pi f\cdot\frac{1}{n}\right) \\ \cos\left(2\pi f\cdot\frac{2}{n}\right) \\ \vdots \\ \cos\left(2\pi f\cdot\frac{n-1}{n}\right) \end{pmatrix} \quad \text{and} \quad \sin 2\pi f\mathbf{t} = \begin{pmatrix} \sin\left(2\pi f\cdot\frac{0}{n}\right) \\ \sin\left(2\pi f\cdot\frac{1}{n}\right) \\ \sin\left(2\pi f\cdot\frac{2}{n}\right) \\ \vdots \\ \sin\left(2\pi f\cdot\frac{n-1}{n}\right) \end{pmatrix},$$

where $\mathbf{t} = (\,0/n \quad 1/n \quad 2/n \quad \cdots \quad n{-}1/n\,)^T$ is the ***discrete time vector***. If the ***discrete exponential*** vectors $\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}}$ and $\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}}$ are defined in the natural way as $\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} = \cos 2\pi f\mathbf{t} + \mathrm{i}\sin 2\pi f\mathbf{t}$ and $\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}} = \cos 2\pi f\mathbf{t} - \mathrm{i}\sin 2\pi f\mathbf{t}$, and if $0 \le f < n$ is an integer frequency,[51] then

$$\mathrm{e}^{\mathrm{i}2\pi f\mathbf{t}} = \begin{pmatrix} \omega^{0f} \\ \omega^{1f} \\ \omega^{2f} \\ \vdots \\ \omega^{(n-1)f} \end{pmatrix} = n\left[\mathbf{F}_n^{-1}\right]_{*f} = n\mathbf{F}_n^{-1}\mathbf{e}_f,$$

where $\mathbf{e}_f$ is the $n \times 1$ unit vector with a 1 in the fth component—remember that components of vectors are indexed from 0 to $n-1$ throughout this section. Similarly, the fact that

$$\xi^{kf} = \omega^{-kf} = 1\omega^{-kf} = \omega^{kn}\omega^{-kf} = \omega^{k(n-f)} \quad \text{for} \quad k = 0, 1, 2, \ldots$$

allows us to conclude that if $0 \le n - f < n$, then

$$\mathrm{e}^{-\mathrm{i}2\pi f\mathbf{t}} = \begin{pmatrix} \xi^{0f} \\ \xi^{1f} \\ \xi^{2f} \\ \vdots \\ \xi^{(n-1)f} \end{pmatrix} = \begin{pmatrix} \omega^{0(n-f)} \\ \omega^{1(n-f)} \\ \omega^{2(n-f)} \\ \vdots \\ \omega^{(n-1)(n-f)} \end{pmatrix} = n\left[\mathbf{F}_n^{-1}\right]_{*n-f} = n\mathbf{F}_n^{-1}\mathbf{e}_{n-f}.$$

[51] The assumption that frequencies are integers is not overly harsh because the Fourier series for a periodic function requires only integer frequencies—recall Example 5.4.6.

Therefore, if $0 < f < n$, then

$$\mathbf{F}_n e^{i2\pi f\mathbf{t}} = n\mathbf{e}_f \quad \text{and} \quad \mathbf{F}_n e^{-i2\pi f\mathbf{t}} = n\mathbf{e}_{n-f}. \tag{5.8.4}$$

Because $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$ and $\sin\theta = (e^{i\theta} - e^{-i\theta})/2i$, it follows from (5.8.4) that for any scalars α and β,

$$\mathbf{F}_n(\alpha\cos 2\pi f\mathbf{t}) = \alpha\mathbf{F}_n\left(\frac{e^{i2\pi f\mathbf{t}} + e^{-i2\pi f\mathbf{t}}}{2}\right) = \frac{n\alpha}{2}\left(\mathbf{e}_f + \mathbf{e}_{n-f}\right)$$

and

$$\mathbf{F}_n(\beta\sin 2\pi f\mathbf{t}) = \beta\mathbf{F}_n\left(\frac{e^{i2\pi f\mathbf{t}} - e^{-i2\pi f\mathbf{t}}}{2i}\right) = \frac{n\beta}{2i}\left(\mathbf{e}_f - \mathbf{e}_{n-f}\right),$$

so that

$$\frac{2}{n}\mathbf{F}_n(\alpha\cos 2\pi f\mathbf{t}) = \alpha\mathbf{e}_f + \alpha\mathbf{e}_{n-f} \tag{5.8.5}$$

and

$$\frac{2}{n}\mathbf{F}_n(\beta\sin 2\pi f\mathbf{t}) = -\beta i\mathbf{e}_f + \beta i\mathbf{e}_{n-f}. \tag{5.8.6}$$

The trigonometric functions $\alpha\cos 2\pi f\tau$ and $\beta\sin 2\pi f\tau$ have amplitudes α and β, respectively, and their frequency is f (their period is $1/f$). The discrete vectors $\alpha\cos 2\pi f\mathbf{t}$ and $\beta\sin 2\pi f\mathbf{t}$ are obtained by evaluating $\alpha\cos 2\pi f\tau$ and $\beta\sin 2\pi f\tau$ at the discrete points in $\mathbf{t} = \begin{pmatrix} 0 & 1/n & 2/n & \cdots & (n-1)/n \end{pmatrix}^T$. As depicted in Figure 5.8.5 for $n = 32$ and $f = 4$, the vectors $\alpha\mathbf{e}_f$ and $\alpha\mathbf{e}_{n-f}$ are interpreted as two pulses of magnitude α at frequencies f and $n - f$.

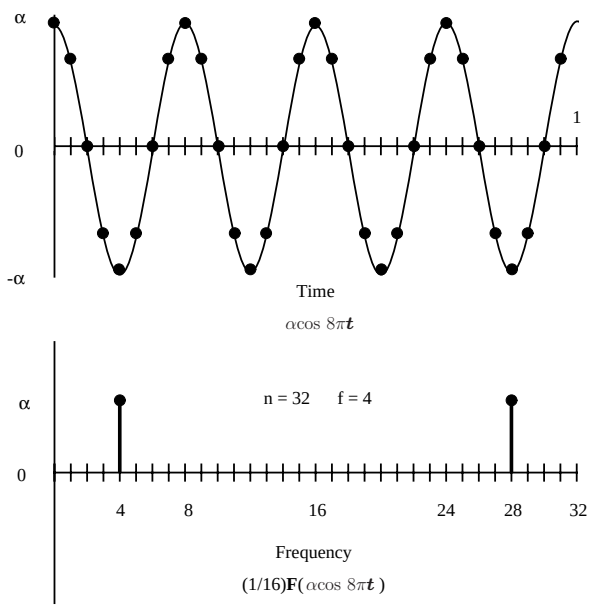


FIGURE 5.8.5

The vector $\alpha \cos 2\pi f\mathbf{t}$ is said to be in the ***time domain***, while the pulses $\alpha\mathbf{e}_f$ and $\alpha\mathbf{e}_{n-f}$ are said to be in the ***frequency domain***. The situation for $\beta \sin 2\pi f\mathbf{t}$ is similarly depicted in Figure 5.8.6 in which $-\beta\mathrm{i}\mathbf{e}_f$ and $\beta\mathrm{i}\mathbf{e}_{n-f}$ are considered two pulses of height $-\beta$ and β, respectively.

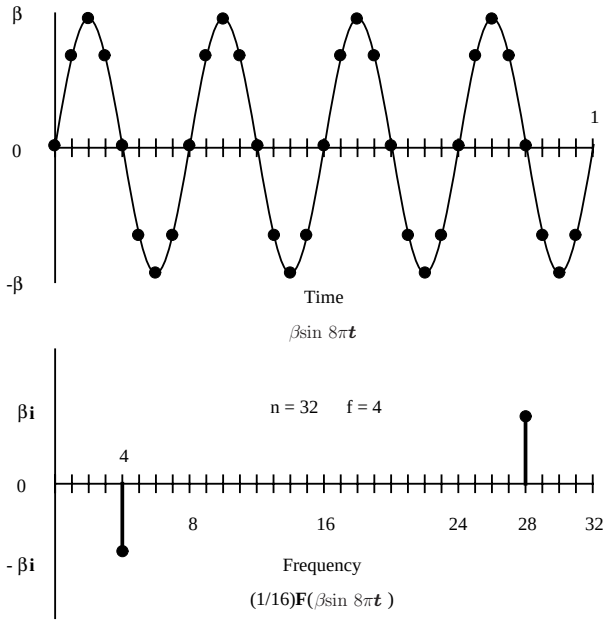


FIGURE 5.8.6

Therefore, if a waveform is given by a finite sum

$$x(\tau) = \sum_k (\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau)$$

in which the f_k's are integers, and if $\mathbf{x}$ is the vector containing the values of $x(\tau)$ at n equally spaced points between time $\tau = 0$ and $\tau = 1$, then, provided that n is sufficiently large,

$$\begin{aligned}\frac{2}{n}\mathbf{F}_n\mathbf{x} &= \frac{2}{n}\mathbf{F}_n\left(\sum_k \alpha_k \cos 2\pi f_k\mathbf{t} + \beta_k \sin 2\pi f_k\mathbf{t}\right)\\ &= \sum_k \frac{2}{n}\mathbf{F}_n\left(\alpha_k \cos 2\pi f_k\mathbf{t}\right) + \sum_k \frac{2}{n}\mathbf{F}_n\left(\beta_k \sin 2\pi f_k\mathbf{t}\right) \qquad (5.8.7)\\ &= \sum_k \alpha_k\left(\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right) + \mathrm{i}\sum_k \beta_k\left(-\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right),\end{aligned}$$

and this exposes the frequency and amplitude of each of the components. If n is chosen so that $\max\{f_k\} < n/2$, then the pulses represented by $\mathbf{e}_f$ and $\mathbf{e}_{n-f}$ are

symmetric about the point $n/2$ in the frequency domain, and the information in just the first (or second) half of the frequency domain completely characterizes the original waveform—this is why only 512/2=256 points are plotted in the graphs shown in Figure 5.8.4. In other words, if

$$\mathbf{y} = \frac{2}{n}\mathbf{F}_n\mathbf{x} = \sum_k \alpha_k \left(\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right) + \mathrm{i}\sum_k \beta_k \left(-\mathbf{e}_{f_k} + \mathbf{e}_{n-f_k}\right), \tag{5.8.8}$$

then the information in

$$\mathbf{y}_{n/2} = \sum_k \alpha_k \mathbf{e}_{f_k} - \mathrm{i}\sum_k \beta_k \mathbf{e}_{f_k} \qquad \text{(the first half of } \mathbf{y}\text{)}$$

is enough to reconstruct the original waveform. For example, the equation of the waveform shown in Figure 5.8.7 is

$$x(\tau) = 3\cos 2\pi\tau + 5\sin 2\pi\tau, \tag{5.8.9}$$

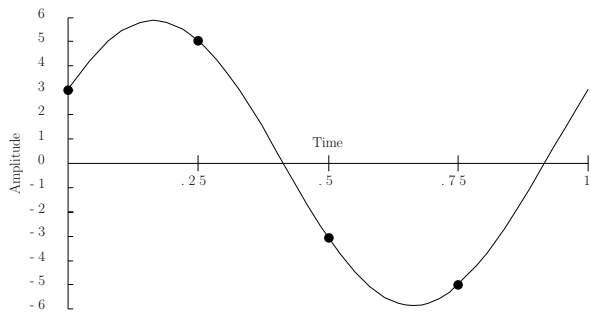


FIGURE 5.8.7

and it is completely determined by the four values in

$$\mathbf{x} = \begin{pmatrix} x(0) \\ x(1/4) \\ x(1/2) \\ x(3/4) \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ -3 \\ -5 \end{pmatrix}.$$

To capture equation (5.8.9) from these four values, compute the vector $\mathbf{y}$ defined by (5.8.8) to be

$$\mathbf{y} = \frac{2}{4}\mathbf{F}_4\mathbf{x} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ -3 \\ -5 \end{pmatrix} = \begin{pmatrix} 0 \\ 3-5\mathrm{i} \\ 0 \\ 3+5\mathrm{i} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 3 \\ 0 \\ 3 \end{pmatrix} + \mathrm{i}\begin{pmatrix} 0 \\ -5 \\ 0 \\ 5 \end{pmatrix} = 3(\mathbf{e}_1 + \mathbf{e}_3) + 5\mathrm{i}(-\mathbf{e}_1 + \mathbf{e}_3).$$

The real part of $\mathbf{y}$ tells us there is a cosine component with $amplitude = 3$ and $frequency = 1$, while the imaginary part of $\mathbf{y}$ says there is a sine component with $amplitude = 5$ and $frequency = 1$. This is depicted in the frequency domain shown in Figure 5.8.8.

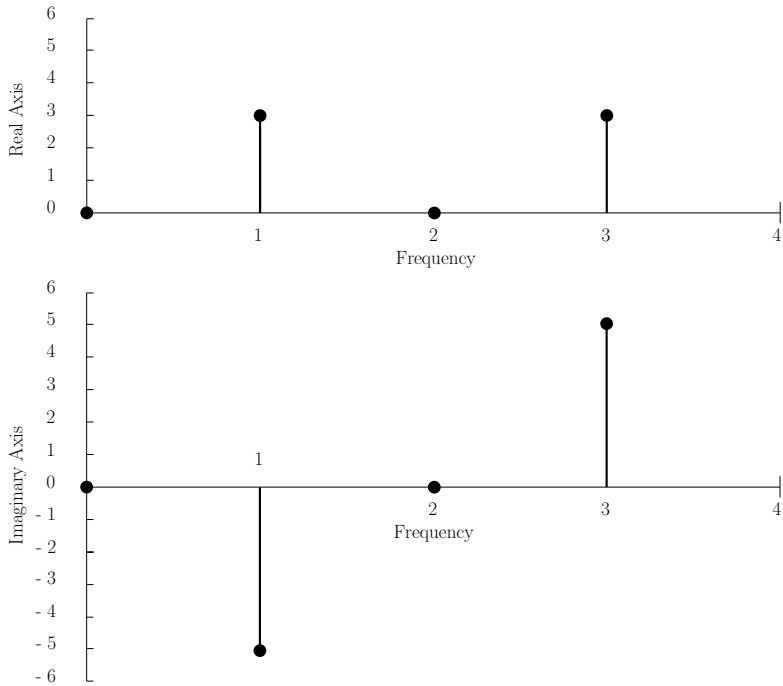


FIGURE 5.8.8

Putting this information together allows us to conclude that the equation of the waveform must be $x(\tau) = 3\cos 2\pi\tau + 5\sin 2\pi\tau$. Since

$$1 = \max\{f_k\} < \frac{n}{2} = \frac{4}{2} = 2,$$

the information in just the first half of $\mathbf{y}$

$$\mathbf{y}_{n/2} = \begin{pmatrix} 0 \\ 3 \end{pmatrix} + \mathrm{i}\begin{pmatrix} 0 \\ -5 \end{pmatrix} = 3\mathbf{e}_1 - 5\mathrm{i}\mathbf{e}_1$$

suffices to completely characterize $x(\tau)$.

These elementary ideas help explain why applying $\mathbf{F}$ to a sample from a signal can reveal the oscillatory components of the signal. But there is still a significant amount of theory that is well beyond the scope of this example. The purpose here is to just hint at how useful the discrete Fourier transform is and why it is so important in analyzing the nature of complicated waveforms.

If

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}_{n\times 1} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix}_{n\times 1},$$

then the vector

$$\mathbf{a} \odot \mathbf{b} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_0\beta_1 + \alpha_1\beta_0 \\ \alpha_0\beta_2 + \alpha_1\beta_1 + \alpha_2\beta_0 \\ \vdots \\ \alpha_{n-2}\beta_{n-1} + \alpha_{n-1}\beta_{n-2} \\ \alpha_{n-1}\beta_{n-1} \\ 0 \end{pmatrix}_{2n\times 1} \tag{5.8.10}$$

is called the ***convolution*** of $\mathbf{a}$ and $\mathbf{b}$. The 0 in the last position is for convenience only—it makes the size of the convolution twice the size of the original vectors, and this provides a balance in some of the formulas involving convolution. Furthermore, it is sometimes convenient to pad $\mathbf{a}$ and $\mathbf{b}$ with n additional zeros to consider them to be vectors with $2n$ components. Setting $\alpha_n = \cdots = \alpha_{2n-1} = \beta_n = \cdots = \beta_{2n-1} = 0$ allows us to write the kth entry in $\mathbf{a} \odot \mathbf{b}$ as

$$[\mathbf{a} \odot \mathbf{b}]_k = \sum_{j=0}^{k} \alpha_j \beta_{k-j} \quad \text{for} \quad k = 0, 1, 2, \ldots, 2n-1.$$

A visual way to form $\mathbf{a} \odot \mathbf{b}$ is to "slide" the reversal of $\mathbf{b}$ "against" $\mathbf{a}$ as depicted in Figure 5.8.9, and then sum the resulting products.

$$\begin{array}{cccccc}
\beta_{n-1} & \beta_{n-1} & \beta_{n-1} & & \alpha_0 & \alpha_0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
\beta_1 & \alpha_0 \times \beta_1 & \alpha_0 \times \beta_2 & & & \alpha_{n-2} \\
\alpha_0 \times \beta_0 & \alpha_1 \times \beta_0 & \alpha_1 \times \beta_1 & \cdots & \alpha_{n-2} \times \beta_{n-1} & \alpha_{n-1} \times \beta_{n-1} \\
\alpha_1 & \vdots & \alpha_2 \times \beta_0 & & \alpha_{n-1} \times \beta_{n-2} & \beta_{n-2} \\
\vdots & \alpha_{n-1} & \vdots & & \vdots & \vdots \\
\alpha_{n-1} & & \alpha_{n-1} & & \beta_0 & \beta_0
\end{array}$$

FIGURE 5.8.9

The convolution operation is a natural occurrence in a variety of situations, and polynomial multiplication is one such example.

Example 5.8.4

Polynomial Multiplication. For $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$, $q(x) = \sum_{k=0}^{n-1} \beta_k x^k$, let $\mathbf{a} = \begin{pmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_{n-1} \end{pmatrix}^T$ and $\mathbf{b} = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_{n-1} \end{pmatrix}^T$. The product $p(x)q(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_{2n-2} x^{2n-2}$ is a polynomial of degree $2n-2$ in which γ_k is simply the kth component of the convolution $\mathbf{a} \odot \mathbf{b}$ because

$$p(x)q(x) = \sum_{k=0}^{2n-2} \left[\sum_{j=0}^{k} \alpha_j \beta_{k-j} \right] x^k = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k x^k. \tag{5.8.11}$$

In other words, polynomial multiplication and convolution are equivalent operations, so if we can devise an efficient way to perform a convolution, then we can efficiently multiply two polynomials, and conversely.

There are two facets involved in efficiently performing a convolution. The first is the realization that the discrete Fourier transform has the ability to convert a convolution into an ordinary product, and vice versa. The second is the realization that it's possible to devise a fast algorithm to compute a discrete Fourier transform. These two facets are developed below.

Convolution Theorem

Let $\mathbf{a} \times \mathbf{b}$ denote the entry-by-entry product

$$\mathbf{a} \times \mathbf{b} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_1\beta_1 \\ \vdots \\ \alpha_{n-1}\beta_{n-1} \end{pmatrix}_{n\times 1},$$

and let $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ be the padded vectors

$$\hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1} \quad \text{and} \quad \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1}.$$

If $\mathbf{F} = \mathbf{F}_{2n}$ is the Fourier matrix of order $2n$, then

$$\mathbf{F}(\mathbf{a} \odot \mathbf{b}) = (\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}}) \quad \text{and} \quad \mathbf{a} \odot \mathbf{b} = \mathbf{F}^{-1}\big[(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}})\big]. \tag{5.8.12}$$

Proof. Observe that the tth component in $\mathbf{F}_{*j} \times \mathbf{F}_{*k}$ is

$$[\mathbf{F}_{*j} \times \mathbf{F}_{*k}]_t = \xi^{tj}\xi^{tk} = \xi^{t(j+k)} = [\mathbf{F}_{*j+k}]_t\,,$$

so that the columns of $\mathbf{F}$ have the property that

$$\mathbf{F}_{*j} \times \mathbf{F}_{*k} = \mathbf{F}_{*j+k} \quad \text{for each} \quad j,k = 0,1,\ldots,(n-1).$$

This means that if $\mathbf{F}\hat{\mathbf{a}}$, $\mathbf{F}\hat{\mathbf{b}}$, and $\mathbf{F}(\mathbf{a} \odot \mathbf{b})$ are expressed as combinations of columns of $\mathbf{F}$ as indicated below,

$$\mathbf{F}\hat{\mathbf{a}} = \sum_{k=0}^{n-1} \alpha_k \mathbf{F}_{*k}, \quad \mathbf{F}\hat{\mathbf{b}} = \sum_{k=0}^{n-1} \beta_k \mathbf{F}_{*k}, \quad \text{and} \quad \mathbf{F}(\mathbf{a} \odot \mathbf{b}) = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k \mathbf{F}_{*k},$$

then the computation of $(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}})$ is exactly the same as forming the product of two polynomials in the sense that

$$\begin{aligned}(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}}) &= \left(\sum_{k=0}^{n-1} \alpha_k \mathbf{F}_{*k}\right)\left(\sum_{k=0}^{n-1} \beta_k \mathbf{F}_{*k}\right) = \sum_{k=0}^{2n-2} \left[\sum_{j=0}^{k} \alpha_j \beta_{k-j}\right] \mathbf{F}_{*k} \\ &= \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{b}]_k \mathbf{F}_{*k} = \mathbf{F}(\mathbf{a} \odot \mathbf{b}). \quad \blacksquare\end{aligned}$$

According to the convolution theorem, the convolution of two $n \times 1$ vectors can be computed by executing three discrete Fourier transforms of order $2n$

$$\mathbf{a}_{n\times 1} \odot \mathbf{b}_{n\times 1} = \mathbf{F}_{2n}^{-1}\big[(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})\big]. \tag{5.8.13}$$

The fact that one of them is an inverse transform is not a source of difficulty—recall Example 5.8.2. But it is still not clear that much has been accomplished. Performing a convolution by following the recipe called for in definition (5.8.10) requires n^2 scalar multiplications (you are asked to verify this in the exercises). Performing a discrete Fourier transform of order $2n$ by standard matrix–vector multiplication requires $4n^2$ scalar multiplications, so using matrix–vector multiplication to perform the computations on the right-hand side of (5.8.13) requires at least 12 times the number of scalar multiplications demanded by the definition of convolution. So, if there is an advantage to be gained by using the convolution theorem, then it is necessary to be able to perform a discrete Fourier transform in far fewer scalar multiplications than that required by standard matrix–vector multiplication. It was not until 1965 that this hurdle was overcome. Two Americans, J. W. Cooley and J. W. Tukey, introduced a fast Fourier transform (FFT) algorithm that requires only on the order of $(n/2)\log_2 n$ scalar multiplications to compute $\mathbf{F}_n\mathbf{x}$. Using the FFT together with the convolution theorem requires

only about $3n\log_2 n$ multiplications to perform a convolution of two $n \times 1$ vectors, and when n is large, this is significantly less than the n^2 factor demanded by the definition of convolution.

The magic of the fast Fourier transform algorithm emanates from the fact that if n is a power of 2, then a discrete Fourier transform of order n can be executed by performing two transforms of order $n/2$. To appreciate exactly how this comes about, observe that when $n = 2^r$ we have $\left(\xi^j\right)^n = \left(\xi^{2j}\right)^{n/2}$, so

$$\left\{1,\, \xi,\, \xi^2,\, \xi^3,\, \ldots,\, \xi^{n-1}\right\} = \text{ the } n^{th} \text{ roots of unity}$$
$$\text{if and only if}$$
$$\left\{1,\, \xi^2,\, \xi^4,\, \xi^6,\, \ldots,\, \xi^{n-2}\right\} = \text{ the } (n/2)^{th} \text{ roots of unity.}$$

This means that the (j,k)-entries in the Fourier matrices $\mathbf{F}_n$ and $\mathbf{F}_{n/2}$ are

$$[\mathbf{F}_n]_{jk} = \xi^{jk} \quad \text{and} \quad [\mathbf{F}_{n/2}]_{jk} = (\xi^2)^{jk} = \xi^{2jk}. \tag{5.8.14}$$

If the columns of $\mathbf{F}_n$ are permuted so that columns with even subscripts are listed before those with odd subscripts, and if $\mathbf{P}_n^T$ is the corresponding permutation matrix, then we can partition $\mathbf{F}_n\mathbf{P}_n^T$ as

$$\mathbf{F}_n\mathbf{P}_n^T = [\mathbf{F}_{*0}\,\mathbf{F}_{*2}\,\cdots\,\mathbf{F}_{*n-2}\,|\,\mathbf{F}_{*1}\,\mathbf{F}_{*3}\,\cdots\,\mathbf{F}_{*n-1}] = \begin{pmatrix} \mathbf{A}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{B}_{\frac{n}{2}\times\frac{n}{2}} \\ \mathbf{C}_{\frac{n}{2}\times\frac{n}{2}} & \mathbf{G}_{\frac{n}{2}\times\frac{n}{2}} \end{pmatrix}.$$

By using (5.8.14) together with the facts that

$$\xi^{nk} = 1 \quad \text{and} \quad \xi^{n/2} = \cos\frac{2\pi(n/2)}{n} - \mathrm{i}\sin\frac{2\pi(n/2)}{n} = -1,$$

we see that the entries in $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ are

$$\mathbf{A}_{jk} = \mathbf{F}_{j,2k} = \xi^{2jk} = [\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{B}_{jk} = \mathbf{F}_{j,2k+1} = \xi^{j(2k+1)} = \xi^j\xi^{2jk} = \xi^j[\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{C}_{jk} = \mathbf{F}_{\frac{n}{2}+j,\,2k} = \xi^{(\frac{n}{2}+j)2k} = \xi^{nk}\xi^{2jk} = \xi^{2jk} = [\mathbf{F}_{n/2}]_{jk},$$

$$\mathbf{G}_{jk} = \mathbf{F}_{\frac{n}{2}+j,\,2k+1} = \xi^{(\frac{n}{2}+j)(2k+1)} = \xi^{nk}\xi^{n/2}\xi^j\xi^{2jk} = -\xi^j\xi^{2jt} = -\xi^j[\mathbf{F}_{n/2}]_{jk}.$$

In other words, if $\mathbf{D}_{n/2}$ is the diagonal matrix

$$\mathbf{D}_{n/2} = \begin{pmatrix} 1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1} \end{pmatrix},$$

then

$$\mathbf{F}_n\mathbf{P}_n^T = \begin{pmatrix} \mathbf{A}_{(n/2)\times(n/2)} & \mathbf{B}_{(n/2)\times(n/2)} \\ \mathbf{C}_{(n/2)\times(n/2)} & \mathbf{G}_{(n/2)\times(n/2)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2} \\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2} \end{pmatrix}.$$

This fundamental feature of the discrete Fourier transform is summarized below.

Decomposing the Fourier Matrix

If $n = 2^r$, then

$$\mathbf{F}_n = \begin{pmatrix} \mathbf{F}_{n/2} & \mathbf{D}_{n/2}\mathbf{F}_{n/2} \\ \mathbf{F}_{n/2} & -\mathbf{D}_{n/2}\mathbf{F}_{n/2} \end{pmatrix} \mathbf{P}_n, \tag{5.8.15}$$

where

$$\mathbf{D}_{n/2} = \begin{pmatrix} 1 & & & & \\ & \xi & & & \\ & & \xi^2 & & \\ & & & \ddots & \\ & & & & \xi^{\frac{n}{2}-1} \end{pmatrix}$$

contains half of the nth roots of unity and $\mathbf{P}_n$ is the "even–odd" permutation matrix defined by

$$\mathbf{P}_n^T = [\mathbf{e}_0\, \mathbf{e}_2\, \mathbf{e}_4\, \cdots\, \mathbf{e}_{n-2} \,|\, \mathbf{e}_1\, \mathbf{e}_3\, \mathbf{e}_5\, \cdots\, \mathbf{e}_{n-1}].$$

The decomposition (5.8.15) says that a discrete Fourier transform of order $n = 2^r$ can be accomplished by two Fourier transforms of order $n/2 = 2^{r-1}$, and this leads to the FFT algorithm. To get a feel for how the FFT works, consider the case when $n = 8$, and proceed to "divide and conquer." If

$$\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}, \quad \text{then} \quad \mathbf{P}_8\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \\ \hline x_1 \\ x_3 \\ x_5 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \hline \mathbf{x}_4^{(1)} \end{pmatrix},$$

so

$$\mathbf{F}_8\mathbf{x}_8 = \begin{pmatrix} \mathbf{F}_4 & \mathbf{D}_4\mathbf{F}_4 \\ \mathbf{F}_4 & -\mathbf{D}_4\mathbf{F}_4 \end{pmatrix} \begin{pmatrix} \mathbf{x}_4^{(0)} \\ \mathbf{x}_4^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_4\mathbf{x}_4^{(0)} + \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \\ \mathbf{F}_4\mathbf{x}_4^{(0)} - \mathbf{D}_4\mathbf{F}_4\mathbf{x}_4^{(1)} \end{pmatrix}. \tag{5.8.16}$$

But

$$\mathbf{P}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \end{pmatrix} \quad \text{and} \quad \mathbf{P}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix},$$

so

$$\mathbf{F}_4\mathbf{x}_4^{(0)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \mathbf{x}_2^{(1)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(0)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \\ \mathbf{F}_2\mathbf{x}_2^{(0)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(1)} \end{pmatrix}$$

and (5.8.17)

$$\mathbf{F}_4\mathbf{x}_4^{(1)} = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix} \begin{pmatrix} \mathbf{x}_2^{(2)} \\ \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} \mathbf{F}_2\mathbf{x}_2^{(2)} + \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \\ \mathbf{F}_2\mathbf{x}_2^{(2)} - \mathbf{D}_2\mathbf{F}_2\mathbf{x}_2^{(3)} \end{pmatrix}.$$

Now, since $\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, it is a trivial matter to compute the terms

$$\mathbf{F}_2\mathbf{x}_2^{(0)}, \quad \mathbf{F}_2\mathbf{x}_2^{(1)}, \quad \mathbf{F}_2\mathbf{x}_2^{(2)}, \quad \mathbf{F}_2\mathbf{x}_2^{(3)}.$$

Of course, to actually carry out the computation, we need to work backward through the preceding sequence of steps. That is, we start with

$$\tilde{\mathbf{x}}_8 = \begin{pmatrix} \mathbf{x}_2^{(0)} \\ \hline \mathbf{x}_2^{(1)} \\ \hline \mathbf{x}_2^{(2)} \\ \hline \mathbf{x}_2^{(3)} \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ \hline x_2 \\ x_6 \\ \hline x_1 \\ x_5 \\ \hline x_3 \\ x_7 \end{pmatrix}, \tag{5.8.18}$$

and use (5.8.17) followed by (5.8.16) to work downward in the following tree.

$$\begin{tikzcd}[column sep=small]
\mathbf{F}_2\mathbf{x}_2^{(0)} \arrow[dr] & & \mathbf{F}_2\mathbf{x}_2^{(1)} \arrow[dl] & & \mathbf{F}_2\mathbf{x}_2^{(2)} \arrow[dr] & & \mathbf{F}_2\mathbf{x}_2^{(3)} \arrow[dl] \\
& \mathbf{F}_4\mathbf{x}_4^{(0)} \arrow[dr] & & & & \mathbf{F}_4\mathbf{x}_4^{(1)} \arrow[dl] & \\
& & \mathbf{F}_8\mathbf{x}_8 & & & &
\end{tikzcd}$$

But there appears to be a snag. In order to work downward through this tree, we cannot start directly with $\mathbf{x}_8$—we must start with the permutation $\tilde{\mathbf{x}}_8$ shown in (5.8.18). So how is this initial permutation determined? Looking back reveals that the entries in $\tilde{\mathbf{x}}_8$ were obtained by first sorting the x_j's into two groups—the entries in the even positions were separated from those in the odd

positions. Then each group was broken into two more groups by again separating the entries in the even positions from those in the odd positions.

$$\begin{array}{ccccccccc}
 & & & & \begin{pmatrix}0 & 1 & 2 & 3 & 4 & 5 & 6 & 7\end{pmatrix} & & & \\
 & & & & \swarrow & \searrow & & & \\
 & & & \begin{pmatrix}0 & 2 & 4 & 6\end{pmatrix} & & \begin{pmatrix}1 & 3 & 5 & 7\end{pmatrix} & & \\
 & & \swarrow & \searrow & & & \swarrow & \searrow & \\
 & \begin{pmatrix}0 & 4\end{pmatrix} & & \begin{pmatrix}2 & 6\end{pmatrix} & & \begin{pmatrix}1 & 5\end{pmatrix} & & \begin{pmatrix}3 & 7\end{pmatrix}
\end{array} \tag{5.8.19}$$

In general, this even–odd sorting process (sometimes called a ***perfect shuffle***) produces the permutation necessary to initiate the algorithm. A clever way to perform a perfect shuffle is to use binary representations and observe that the first level of sorting in (5.8.19) is determined according to whether the least significant bit is 0 or 1, the second level of sorting is determined by the second least significant bit, and so on—this is illustrated in Table 5.8.1 for $n = 8$.

Table 5.8.1

Natural order	**First level**	**Second level**
$0 \leftrightarrow 000$	$0 \leftrightarrow 00\mathbf{0}$	$0 \leftrightarrow 0\mathbf{0}0$
$1 \leftrightarrow 001$	$2 \leftrightarrow 01\mathbf{0}$	$4 \leftrightarrow 1\mathbf{0}0$
$2 \leftrightarrow 010$	$4 \leftrightarrow 10\mathbf{0}$	$2 \leftrightarrow 0\mathbf{1}0$
$3 \leftrightarrow 011$	$6 \leftrightarrow 11\mathbf{0}$	$6 \leftrightarrow 1\mathbf{1}0$
$4 \leftrightarrow 100$	$1 \leftrightarrow 00\mathbf{1}$	$1 \leftrightarrow 0\mathbf{0}1$
$5 \leftrightarrow 101$	$3 \leftrightarrow 01\mathbf{1}$	$5 \leftrightarrow 1\mathbf{0}1$
$6 \leftrightarrow 110$	$5 \leftrightarrow 10\mathbf{1}$	$3 \leftrightarrow 0\mathbf{1}1$
$7 \leftrightarrow 111$	$7 \leftrightarrow 11\mathbf{1}$	$7 \leftrightarrow 1\mathbf{1}1$

But all intermediate levels in this sorting process can be eliminated because something very nice occurs. Examination of the last column in Table 5.8.1 reveals that the binary bits in the perfect shuffle ordering are exactly the reversal of the binary bits in the natural ordering. In other words,

- to generate the perfect shuffle of the numbers $0, 1, 2, \ldots, n-1$, simply reverse the bits in the binary representation of each number.

We can summarize the fast Fourier transform by the following implementation that utilizes array operations.[52]

[52] There are a variety of different ways to implement the FFT, and choosing a practical implementation frequently depends on the hardware being used as well as the application under consideration. The FFT ranks high on the list of useful algorithms because it provides an advantage in a large variety of applications, and there are many more facets of the FFT than those presented here (e.g., FFT when n is not a power of 2). In fact, there are entire texts devoted to these issues, so the interested student need only go as far as the nearest library to find more details.

Fast Fourier Transform

For a given input vector $\mathbf{x}$ containing $n = 2^r$ components, the discrete Fourier transform $\mathbf{F}_n\mathbf{x}$ is the result of successively creating the following arrays.

$\mathbf{X}_{1\times n} \longleftarrow rev(\mathbf{x})$ (bit reverse the subscripts)

For $j = 0, 1, 2, 3, \ldots, r-1$

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ e^{-\pi i/2^j} \\ e^{-2\pi i/2^j} \\ e^{-3\pi i/2^j} \\ \vdots \\ e^{-(2^j-1)\pi i/2^j} \end{pmatrix}_{2^j\times 1} \quad \text{(Half of the } (2^{j+1})^{th} \text{ roots of 1, perhaps from a lookup table)}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*0} & \mathbf{X}_{*2} & \mathbf{X}_{*4} & \cdots & \mathbf{X}_{*2^{r-j}-2} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} \mathbf{X}_{*1} & \mathbf{X}_{*3} & \mathbf{X}_{*5} & \cdots & \mathbf{X}_{*2^{r-j}-1} \end{pmatrix}_{2^j\times 2^{r-j-1}}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D}\times\mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D}\times\mathbf{X}^{(1)} \end{pmatrix}_{2^{j+1}\times 2^{r-j-1}} \quad \left(\begin{array}{l}\text{Define } \times \text{ to mean} \\ [\mathbf{D}\times\mathbf{M}]_{ij} = d_i m_{ij}\end{array}\right)$$

Example 5.8.5

Problem: Perform the FFT on $\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

Solution: Start with $\mathbf{X} \longleftarrow rev(\mathbf{x}) = (\, x_0 \quad x_2 \quad x_1 \quad x_3 \,)$.

For $j = 0$:

$\mathbf{D} \longleftarrow (1)$ (Half of the square roots of 1)

$\mathbf{X}^{(0)} \longleftarrow (\, x_0 \quad x_1 \,)$

$\mathbf{X}^{(1)} \longleftarrow (\, x_2 \quad x_3 \,)$ and $\mathbf{D}\times\mathbf{X}^{(1)} \longleftarrow (\, x_2 \quad x_3 \,)$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D}\times\mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D}\times\mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 & x_1 + x_3 \\ x_0 - x_2 & x_1 - x_3 \end{pmatrix}$$

For $j = 1$:

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ -\mathrm{i} \end{pmatrix} \qquad \text{(Half of the } 4^{th} \text{ roots of 1)}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 + x_2 \\ x_0 - x_2 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_3 \\ x_1 - x_3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} \times \mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_3 \\ -\mathrm{i}x_1 + \mathrm{i}x_3 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix} = \begin{pmatrix} x_0 + x_2 + x_1 + x_3 \\ x_0 - x_2 - \mathrm{i}x_1 + \mathrm{i}x_3 \\ x_0 + x_2 - x_1 - x_3 \\ x_0 - x_2 + \mathrm{i}x_1 - \mathrm{i}x_3 \end{pmatrix} = \mathbf{F}_4\mathbf{x}$$

Notice that this agrees with the result obtained by using direct matrix–vector multiplication with $\mathbf{F}_4$ given in Example 5.8.1.

To understand why it is called the "fast" Fourier transform, simply count the number of multiplications the FFT requires. Observe that the jth iteration requires 2^j multiplications for each column in $\mathbf{X}^{(1)}$, and there are 2^{r-j-1} columns, so 2^{r-1} multiplications are used for each iteration.[53] Since r iterations are required, the total number of multiplications used by the FFT does not exceed $2^{r-1}r = (n/2)\log_2 n$.

FFT Multiplication Count

If n is a power of 2, then applying the FFT to a vector of n components requires at most $(n/2)\log_2 n$ multiplications.

The $(n/2)\log_2 n$ count represents a tremendous advantage over the n^2 factor demanded by a direct matrix–vector product. To appreciate the magnitude of the difference between n^2 and $(n/2)\log_2 n$, look at Figure 5.8.10.

[53] Actually, we can get by with slightly fewer multiplications if we take advantage of the fact that the first entry in $\mathbf{D}$ is always 1 and if we observe that no multiplications are necessary when $j = 0$. But when n is large, these savings are relatively insignificant, so they are ignored in the multiplication count.

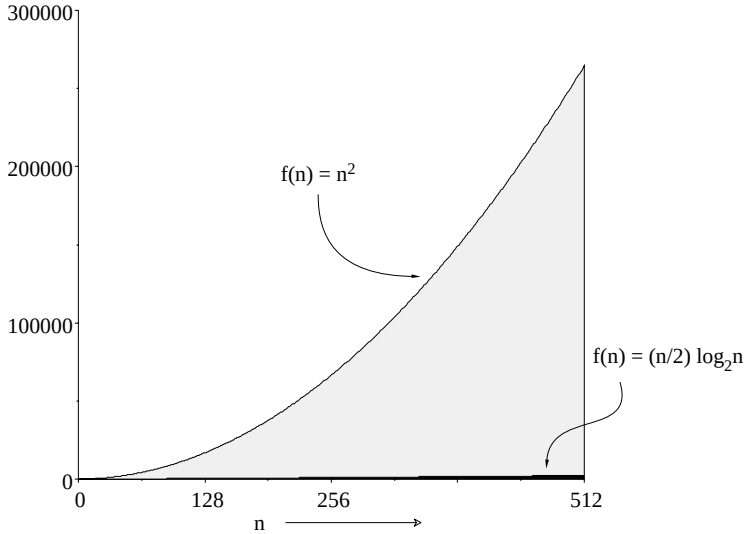


FIGURE 5.8.10

The small dark portion at the bottom of the graph is the area under the curve $f(n) = (n/2)\log_2 n$—it is tiny in comparison to the area under $f(n) = n^2$. For example, if $n = 512$, then $n^2 = 262,144$, but $(n/2)\log_2 n = 2304$. In other words, for $n = 512$, the FFT is on the order of 100 times faster than straightforward matrix–vector multiplication, and for larger values of n, the gap is even wider—Figure 5.8.10 illustrates just how fast the difference between n^2 and $(n/2)\log_2 n$ grows as n increases. Since Cooley and Tukey introduced the FFT in 1965, it has risen to a position of fundamental importance. The FFT and the convolution theorem are extremely powerful tools, and they have been principal components of the computational revolution that now touches our lives countless times each day.

Example 5.8.6

Problem: Fast Integer Multiplication. Consider two positive integers whose base-b representations are

$$c = (\gamma_{n-1}\gamma_{n-2}\cdots\gamma_1\gamma_0)_b \quad \text{and} \quad d = (\delta_{n-1}\delta_{n-2}\cdots\delta_1\delta_0)_b.$$

Use the convolution theorem together with the FFT to compute the product cd.

Solution: If we let

$$p(x) = \sum_{k=0}^{n-1} \gamma_k x^k, \quad q(x) = \sum_{k=0}^{n-1} \delta_k x^k, \quad \mathbf{c} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{n-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{d} = \begin{pmatrix} \delta_0 \\ \delta_1 \\ \vdots \\ \delta_{n-1} \end{pmatrix},$$

then

$$c = \gamma_{n-1}b^{n-1} + \gamma_{n-2}b^{n-2} + \cdots + \gamma_1 b^1 + \gamma_0 b^0 = p(b),$$
$$d = \delta_{n-1}b^{n-1} + \delta_{n-2}b^{n-2} + \cdots + \delta_1 b^1 + \delta_0 b^0 = q(b),$$

and it follows from (5.8.11) that the product cd is given by

$$cd = p(b)q(b) = [\mathbf{c}\odot\mathbf{d}]_{2n-2}b^{2n-2} + [\mathbf{c}\odot\mathbf{d}]_{2n-3}b^{2n-3} + \cdots + [\mathbf{c}\odot\mathbf{d}]_1 b^1 + [\mathbf{c}\odot\mathbf{d}]_0 b^0.$$

It looks as though the convolution $\mathbf{c}\odot\mathbf{d}$ provides the base-b representation for cd, but this is not quite the case because it is possible to have some $[\mathbf{c}\odot\mathbf{d}]_k \geq b$. For example, if $c = 201_{10}$ and $d = 425_{10}$, then

$$\mathbf{c}\odot\mathbf{d} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} \odot \begin{pmatrix} 5 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 14 \\ 4 \\ 8 \\ 0 \end{pmatrix},$$

so

$$cd = (8\times10^4) + (4\times10^3) + (14\times10^2) + (2\times10^1) + (5\times10^0). \tag{5.8.20}$$

But when numbers like 14 (i.e., greater than or equal to the base) appear in $\mathbf{c}\odot\mathbf{d}$, it is relatively easy to decompose them by writing $14 = (1\times10^1) + (4\times10^0)$, so

$$14\times10^2 = \big[(1\times10^1) + (4\times10^0)\big]\times10^2 = (1\times10^3) + (4\times10^2).$$

Substituting this in (5.8.20) and combining coefficients of like powers produces the base-10 representation of the product

$$cd = (8\times10^4) + (5\times10^3) + (4\times10^2) + (2\times10^1) + (5\times10^0) = 85425_{10}.$$

Computing $\mathbf{c}\odot\mathbf{d}$ directly demands n^2 multiplications, but using the FFT in conjunction with the convolution theorem requires only about $3n\log_2 n$ multiplications, which is considerably less than n^2 for large values of n. Thus it is possible to multiply very long base-b integers much faster than by using direct methods. Most digital computers have binary integer multiplication (usually 64-bit multiplication not requiring the FFT) built into their hardware, but for ultra-high-precision multiplication or for more general base-b multiplication, the FFT is a viable tool.

Exercises for section 5.8

5.8.1. Evaluate the following convolutions.

(a) $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \odot \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$. (b) $\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$. (c) $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \odot \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$.

5.8.2. (a) Evaluate the discrete Fourier transform of $\begin{pmatrix} 1 \\ -\mathrm{i} \\ -1 \\ \mathrm{i} \end{pmatrix}$.

(b) Evaluate the inverse transform of $\begin{pmatrix} 1 \\ \mathrm{i} \\ -1 \\ -\mathrm{i} \end{pmatrix}$.

5.8.3. Verify directly that $\mathbf{F}_4 = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix}\mathbf{P}_4$, where the $\mathbf{F}_4$, $\mathbf{P}_4$, and $\mathbf{D}_2$, are as defined in (5.8.15).

5.8.4. Use the following vectors to perform the indicated computations:

$$\mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ 0 \\ 0 \end{pmatrix}.$$

(a) Compute $\mathbf{a} \odot \mathbf{b}$, $\mathbf{F}_4(\mathbf{a} \odot \mathbf{b})$, and $(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})$.

(b) By using $\mathbf{F}_4^{-1}$ as given in Example 5.8.1, compute

$$\mathbf{F}_4^{-1}\big[(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})\big].$$

Compare this with the results guaranteed by the convolution theorem.

5.8.5. For $p(x) = 2x - 3$ and $q(x) = 3x - 4$, compute the product $p(x)q(x)$ by using the convolution theorem.

5.8.6. Use convolutions to form the following products.

(a) $43_{10} \times 21_{10}$. (b) $123_8 \times 601_8$. (c) $1010_2 \times 1101_2$.

5.8.7. Let $\mathbf{a}$ and $\mathbf{b}$ be $n \times 1$ vectors, where n is a power of 2.

(a) Show that the number of multiplications required to form $\mathbf{a} \odot \mathbf{b}$ by using the definition of convolution is n^2.
Hint: $1 + 2 + \cdots + k = k(k+1)/2$.

(b) Show that the number of multiplications required to form $\mathbf{a} \odot \mathbf{b}$ by using the FFT in conjunction with the convolution theorem is $3n \log_2 n + 7n$. Sketch a graph of $3n \log_2 n$ (the $7n$ factor is dropped because it is not significant), and compare it with the graph of n^2 to illustrate why the FFT in conjunction with the convolution theorem provides such a big advantage.

5.8.8. A waveform given by a finite sum

$$x(\tau) = \sum_k (\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau)$$

in which the f_k's are integers and $\max\{f_k\} \le 3$ is sampled at eight equally spaced points between $\tau = 0$ and $\tau = 1$. Let

$$\mathbf{x} = \begin{pmatrix} x(0/8) \\ x(1/8) \\ x(2/8) \\ x(3/8) \\ x(4/8) \\ x(5/8) \\ x(6/8) \\ x(7/8) \end{pmatrix}, \quad \text{and suppose that} \quad \mathbf{y} = \frac{1}{4}\mathbf{F}_8\mathbf{x} = \begin{pmatrix} 0 \\ -5\mathrm{i} \\ 1-3\mathrm{i} \\ 4 \\ 0 \\ 4 \\ 1+3\mathrm{i} \\ 5\mathrm{i} \end{pmatrix}.$$

What is the equation of the waveform?

5.8.9. Prove that $\mathbf{a} \odot \mathbf{b} = \mathbf{b} \odot \mathbf{a}$ for all $\mathbf{a}, \mathbf{b} \in \mathcal{C}^n$—i.e., convolution is a commutative operation.

5.8.10. For $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$ and the nth roots of unity ξ_k, let

$$\mathbf{a} = (\,\alpha_0\ \alpha_1\ \alpha_2\ \cdots \alpha_{n-1}\,)^T \quad \text{and} \quad \mathbf{p} = (\,p(1)\ p(\xi)\ p(\xi^2)\ \cdots p(\xi^{n-1})\,)^T.$$

Explain why $\mathbf{F}_n\mathbf{a} = \mathbf{p}$ and $\mathbf{a} = \mathbf{F}_n^{-1}\mathbf{p}$. This says that the discrete Fourier transform allows us to go from the representation of a polynomial p in terms of its coefficients α_k to the representation of p in terms of its values $p(\xi^k)$, and the inverse transform takes us in the other direction.

5.8.11. For two polynomials $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$ and $q(x) = \sum_{k=0}^{n-1} \beta_k x^k$, let

$$\mathbf{p} = \begin{pmatrix} p(1) \\ p(\xi) \\ \vdots \\ p(\xi^{2n-1}) \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} q(1) \\ q(\xi) \\ \vdots \\ q(\xi^{2n-1}) \end{pmatrix},$$

where $\{1, \xi, \xi^2, \ldots, \xi^{2n-1}\}$ are now the $2n^{th}$ roots of unity. Explain why the coefficients in the product

$$p(x)q(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_{2n-2} x^{2n-2}$$

must be given by

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \end{pmatrix} = \mathbf{F}_{2n}^{-1} \begin{pmatrix} p(1)q(1) \\ p(\xi)q(\xi) \\ p(\xi^2)q(\xi^2) \\ \vdots \end{pmatrix}.$$

This says that the product $p(x)q(x)$ is completely determined by the values of $p(x)$ and $q(x)$ at the $2n^{th}$ roots of unity.

5.8.12. A ***circulant*** matrix is defined to be a square matrix that has the form

$$\mathbf{C} = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}_{n\times n}.$$

In other words, the entries in each column are the same as the previous column, but they are shifted one position downward and wrapped around at the top—the (j,k)-entry in $\mathbf{C}$ can be described as $c_{jk} = c_{j-k \,(\mathrm{mod}\ n)}$. (Some authors use $\mathbf{C}^T$ rather than $\mathbf{C}$ as the definition—it doesn't matter.)

(a) If $\mathbf{Q}$ is the circulant matrix defined by

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{n\times n},$$

and if $p(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$, verify that

$$\mathbf{C} = p(\mathbf{Q}) = c_0\mathbf{I} + c_1\mathbf{Q} + \cdots + c_{n-1}\mathbf{Q}^{n-1}.$$

(b) Explain why the Fourier matrix of order n diagonalizes $\mathbf{Q}$ in the sense that

$$\mathbf{FQF}^{-1} = \mathbf{D} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \xi & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \xi^{n-1} \end{pmatrix},$$

where the ξ^k's are the nth roots of unity.

(c) Prove that the Fourier matrix of order n diagonalizes every $n \times n$ circulant in the sense that

$$\mathbf{FCF}^{-1} = \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix},$$

where $p(x) = c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$.

(d) If $\mathbf{C}_1$ and $\mathbf{C}_2$ are any pair of $n \times n$ circulants, explain why $\mathbf{C}_1\mathbf{C}_2 = \mathbf{C}_2\mathbf{C}_1$—i.e., all circulants commute with each other.

5.8.13. For a nonsingular circulant $\mathbf{C}_{n\times n}$, explain how to use the FFT algorithm to efficiently perform the following operations.

(a) Solve a system $\mathbf{Cx} = \mathbf{b}$.

(b) Compute $\mathbf{C}^{-1}$.

(c) Multiply two circulants $\mathbf{C}_1\mathbf{C}_2$.

5.8.14. For the vectors

$$\mathbf{a}=\begin{pmatrix}\alpha_0\\ \alpha_1\\ \vdots\\ \alpha_{n-1}\end{pmatrix},\quad \mathbf{b}=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_{n-1}\end{pmatrix},\quad \hat{\mathbf{a}}=\begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_{n-1}\\ 0\\ \vdots\\ 0\end{pmatrix}_{2n\times 1},\ \text{and}\ \hat{\mathbf{b}}=\begin{pmatrix}\beta_0\\ \vdots\\ \beta_{n-1}\\ 0\\ \vdots\\ 0\end{pmatrix}_{2n\times 1},$$

let $\mathbf{C}$ be the $2n \times 2n$ circulant matrix (see Exercise 5.8.12) whose first column is $\hat{\mathbf{a}}$.

(a) Show that the convolution operation can be described as a matrix–vector product by demonstrating that

$$\mathbf{a} \odot \mathbf{b} = \mathbf{C}\hat{\mathbf{b}}.$$

(b) Use this relationship to give an alternate proof of the convolution theorem. **Hint:** Use the diagonalization result of Exercise 5.8.12 together with the result of Exercise 5.8.10.

5.8.15. The ***Kronecker product*** of two matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is defined to be the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B}\\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B}\end{pmatrix}.$$

This is also known as the ***tensor product*** or the ***direct product***. Although there is an extensive list of properties that the tensor product satisfies, this exercise requires only the following two elementary facts (which you need not prove unless you feel up to it). The complete list of properties is given in Exercise 7.8.11 (p. 597) along with remarks about Kronecker, and another application appears in Exercise 7.6.10 (p. 573).

$\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$.

$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ (if $\mathbf{AC}$ and $\mathbf{BD}$ exist).

(a) If $n = 2^r$, and if $\mathbf{P}_n$ is the even–odd permutation matrix described in (5.8.15), explain why

$$\mathbf{R}_n = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_{2^1})(\mathbf{I}_{2^{r-2}} \otimes \mathbf{P}_{2^2}) \cdots (\mathbf{I}_{2^1} \otimes \mathbf{P}_{2^{r-1}})(\mathbf{I}_{2^0} \otimes \mathbf{P}_{2^r})$$

is the permutation matrix associated with the bit reversing (or perfect shuffle) permutation described in (5.8.19) and Table 5.8.1. **Hint:** Work it out for $n = 8$ by showing

$$\mathbf{R}_8 \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ x_2 \\ x_6 \\ x_1 \\ x_5 \\ x_3 \\ x_7 \end{pmatrix},$$

and you will see why it holds in general.

(b) Suppose $n = 2^r$, and set

$$\mathbf{B}_n = \begin{pmatrix} \mathbf{I}_{n/2} & \mathbf{D}_{n/2} \\ \mathbf{I}_{n/2} & -\mathbf{D}_{n/2} \end{pmatrix}.$$

According to (5.8.15), the Fourier matrix can be written as

$$\mathbf{F}_n = \mathbf{B}_n(\mathbf{I}_2 \otimes \mathbf{F}_{n/2})\mathbf{P}_n.$$

Expand on this idea by proving that $\mathbf{F}_n$ can be factored as

$$\mathbf{F}_n = \mathbf{L}_n\mathbf{R}_n$$

in which

$$\mathbf{L}_n = (\mathbf{I}_{2^0} \otimes \mathbf{B}_{2^r})(\mathbf{I}_{2^1} \otimes \mathbf{B}_{2^{r-1}}) \cdots (\mathbf{I}_{2^{r-2}} \otimes \mathbf{B}_{2^2})(\mathbf{I}_{2^{r-1}} \otimes \mathbf{B}_{2^1}),$$

and where $\mathbf{R}_n$ is the bit reversing permutation

$$\mathbf{R}_n = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_{2^1})(\mathbf{I}_{2^{r-2}} \otimes \mathbf{P}_{2^2}) \cdots (\mathbf{I}_{2^1} \otimes \mathbf{P}_{2^{r-1}})(\mathbf{I}_{2^0} \otimes \mathbf{P}_{2^r}).$$

Notice that this says $\mathbf{F}_n\mathbf{x} = \mathbf{L}_n\mathbf{R}_n\mathbf{x}$, so the discrete Fourier transform of $\mathbf{x}$ is obtained by first performing the bit reversing permutation to $\mathbf{x}$ followed by r applications of the terms $(\mathbf{I}_{2^{r-k}} \otimes \mathbf{B}_{2^k})$ from $\mathbf{L}_n$. This in fact is the FFT algorithm in factored form. **Hint:** Define two sequences by the rules

$$\mathbf{L}_{2^k} = (\mathbf{I}_{2^{r-k}} \otimes \mathbf{B}_{2^k})\,\mathbf{L}_{2^{k-1}} \quad \text{and} \quad \mathbf{R}_{2^k} = \mathbf{R}_{2^{k-1}}\,(\mathbf{I}_{2^{r-k}} \otimes \mathbf{P}_{2^k}),$$

where

$$\mathbf{L}_1 = 1, \quad \mathbf{R}_1 = \mathbf{I}_n, \quad \mathbf{B}_2 = \mathbf{F}_2, \quad \mathbf{P}_2 = \mathbf{I}_2,$$

and use induction on k to prove that

$$\mathbf{I}_{2^{r-k}} \otimes \mathbf{F}_{2^k} = \mathbf{L}_{2^k}\mathbf{R}_{2^k} \quad \text{for} \quad k = 1, 2, \ldots, r.$$

5.8.16. For $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_{n-1} x^{n-1}$, prove that

$$\frac{1}{n} \sum_{k=0}^{n-1} \left| p(\xi^k) \right|^2 = |\alpha_0|^2 + |\alpha_1|^2 + \cdots + |\alpha_{n-1}|^2,$$

where $\left\{1, \xi, \xi^2, \ldots, \xi^{n-1}\right\}$ are the nth roots of unity.

5.8.17. Consider a waveform that is given by the finite sum

$$x(\tau) = \sum_k \left(\alpha_k \cos 2\pi f_k \tau + \beta_k \sin 2\pi f_k \tau\right)$$

in which the f_k 's are distinct integers, and let

$$\mathbf{x} = \sum_k \left(\alpha_k \cos 2\pi f_k \mathbf{t} + \beta_k \sin 2\pi f_k \mathbf{t}\right)$$

be the vector containing the values of $x(\tau)$ at $n > 2\max\{f_k\}$ equally spaced points between $\tau = 0$ and $\tau = 1$ as described in Example 5.8.3. Use the discrete Fourier transform to prove that

$$\|\mathbf{x}\|_2^2 = \frac{n}{2} \sum_k \left(\alpha_k^2 + \beta_k^2\right).$$

5.8.18. Let η be an arbitrary scalar, and let

$$\mathbf{c} = \begin{pmatrix} 1 \\ \eta \\ \eta^2 \\ \vdots \\ \eta^{2n-1} \end{pmatrix} \quad \text{and} \quad \mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix}.$$

Prove that $\mathbf{c}^T(\mathbf{a} \odot \mathbf{a}) = \left(\mathbf{c}^T \hat{\mathbf{a}}\right)^2$.

5.8.19. Apply the FFT algorithm to the vector $\mathbf{x}_8 = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_7 \end{pmatrix}$, and then verify that your answer agrees with the result obtained by computing $\mathbf{F}_8 \mathbf{x}_8$ directly.

5.9 COMPLEMENTARY SUBSPACES

The sum of two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of a vector space $\mathcal{V}$ was defined on p. 166 to be the set $\mathcal{X}+\mathcal{Y}=\{\mathbf{x}+\mathbf{y} \mid \mathbf{x}\in\mathcal{X} \text{ and } \mathbf{y}\in\mathcal{Y}\}$, and it was established that $\mathcal{X}+\mathcal{Y}$ is another subspace of $\mathcal{V}$. For example, consider the two subspaces of $\Re^3$ shown in Figure 5.9.1 in which $\mathcal{X}$ is a plane through the origin, and $\mathcal{Y}$ is a line through the origin.

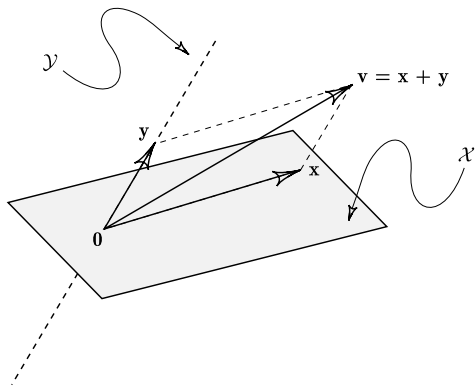


FIGURE 5.9.1

Notice that $\mathcal{X}$ and $\mathcal{Y}$ are ***disjoint*** in the sense that $\mathcal{X}\cap\mathcal{Y}=\mathbf{0}$. The parallelogram law for vector addition makes it clear that $\mathcal{X}+\mathcal{Y}=\Re^3$ because each vector in $\Re^3$ can be written as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$." Thus $\Re^3$ is resolved into a pair of disjoint components $\mathcal{X}$ and $\mathcal{Y}$. These ideas generalize as described below.

Complementary Subspaces

Subspaces $\mathcal{X}$, $\mathcal{Y}$ of a space $\mathcal{V}$ are said to be ***complementary*** whenever

$$\mathcal{V}=\mathcal{X}+\mathcal{Y} \quad \text{and} \quad \mathcal{X}\cap\mathcal{Y}=\mathbf{0}, \tag{5.9.1}$$

in which case $\mathcal{V}$ is said to be the ***direct sum*** of $\mathcal{X}$ and $\mathcal{Y}$, and this is denoted by writing $\mathcal{V}=\mathcal{X}\oplus\mathcal{Y}$.

- For a vector space $\mathcal{V}$ with subspaces $\mathcal{X},\mathcal{Y}$ having respective bases $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$, the following statements are equivalent.
  - ▷ $\mathcal{V}=\mathcal{X}\oplus\mathcal{Y}$. (5.9.2)
  - ▷ For each $\mathbf{v}\in\mathcal{V}$ there are *unique* vectors $\mathbf{x}\in\mathcal{X}$ and $\mathbf{y}\in\mathcal{Y}$ such that $\mathbf{v}=\mathbf{x}+\mathbf{y}$. (5.9.3)
  - ▷ $\mathcal{B}_{\mathcal{X}}\cap\mathcal{B}_{\mathcal{Y}}=\phi$ and $\mathcal{B}_{\mathcal{X}}\cup\mathcal{B}_{\mathcal{Y}}$ is a basis for $\mathcal{V}$. (5.9.4)

Prove these by arguing (5.9.2) $\Longrightarrow$ (5.9.3) $\Longrightarrow$ (5.9.4) $\Longrightarrow$ (5.9.2).

Proof of (5.9.2) $\Longrightarrow$ (5.9.3). First recall from (4.4.19) that

$$\dim \mathcal{V} = \dim(\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim(\mathcal{X} \cap \mathcal{Y}).$$

If $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, then $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$, and thus $\dim \mathcal{V} = \dim \mathcal{X} + \dim \mathcal{Y}$. To prove (5.9.3), suppose there are two ways to represent a vector $\mathbf{v} \in \mathcal{V}$ as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$." If $\mathbf{v} = \mathbf{x}_1 + \mathbf{y}_1 = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then

$$\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{y}_2 - \mathbf{y}_1 \quad \Longrightarrow \quad \left\{ \begin{array}{c} \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{X} \\ \text{and} \\ \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{Y} \end{array} \right\} \quad \Longrightarrow \quad \mathbf{x}_1 - \mathbf{x}_2 \in \mathcal{X} \cap \mathcal{Y}.$$

But $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$, so $\mathbf{x}_1 = \mathbf{x}_2$ and $\mathbf{y}_1 = \mathbf{y}_2$.

Proof of (5.9.3) $\Longrightarrow$ (5.9.4). The hypothesis insures that $\mathcal{V} = \mathcal{X} + \mathcal{Y}$, and we know from (4.1.2) that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ spans $\mathcal{X} + \mathcal{Y}$, so $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ must be a spanning set for $\mathcal{V}$. To prove $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is linearly independent, let $\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_s\}$, and suppose that

$$\mathbf{0} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i + \sum_{j=1}^{s} \beta_j \mathbf{y}_j.$$

This is one way to express $\mathbf{0}$ as "something from $\mathcal{X}$ plus something from $\mathcal{Y}$," while $\mathbf{0} = \mathbf{0} + \mathbf{0}$ is another way. Consequently, (5.9.3) guarantees that

$$\sum_{i=1}^{r} \alpha_i \mathbf{x}_i = \mathbf{0} \quad \text{and} \quad \sum_{j=1}^{s} \beta_j \mathbf{y}_j = \mathbf{0},$$

and hence $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ and $\beta_1 = \beta_2 = \cdots = \beta_s = 0$ because $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are both linearly independent. Therefore, $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is linearly independent, and hence it is a basis for $\mathcal{V}$.

Proof of (5.9.4) $\Longrightarrow$ (5.9.2). If $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\mathcal{V}$, then $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a linearly independent set. This together with the fact that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ always spans $\mathcal{X} + \mathcal{Y}$ means $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\mathcal{X} + \mathcal{Y}$ as well as for $\mathcal{V}$. Consequently, $\mathcal{V} = \mathcal{X} + \mathcal{Y}$, and hence

$$\dim \mathcal{X} + \dim \mathcal{Y} = \dim \mathcal{V} = \dim(\mathcal{X} + \mathcal{Y}) = \dim \mathcal{X} + \dim \mathcal{Y} - \dim(\mathcal{X} \cap \mathcal{Y}),$$

so $\dim(\mathcal{X} \cap \mathcal{Y}) = 0$ or, equivalently, $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$. ∎

If $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, then (5.9.3) says there is one and only one way to resolve each $\mathbf{v} \in \mathcal{V}$ into an "$\mathcal{X}$-component" and a "$\mathcal{Y}$-component" so that $\mathbf{v} = \mathbf{x} + \mathbf{y}$. These two components of $\mathbf{v}$ have a definite geometrical interpretation. Look back at Figure 5.9.1 in which $\Re^3 = \mathcal{X} \oplus \mathcal{Y}$, where $\mathcal{X}$ is a plane and $\mathcal{Y}$ is a line outside the plane, and notice that $\mathbf{x}$ (the $\mathcal{X}$-component of $\mathbf{v}$) is the result of projecting $\mathbf{v}$ onto $\mathcal{X}$ along a line parallel to $\mathcal{Y}$, and $\mathbf{y}$ (the $\mathcal{Y}$-component of $\mathbf{v}$) is obtained by projecting $\mathbf{v}$ onto $\mathcal{Y}$ along a line parallel to $\mathcal{X}$. This leads to the following formal definition of a projection.

Projection

Suppose that $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$ so that for each $\mathbf{v} \in \mathcal{V}$ there are unique vectors $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$.

- The vector $\mathbf{x}$ is called the ***projection*** of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$.
- The vector $\mathbf{y}$ is called the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$.

It's clear that if $\mathcal{X} \perp \mathcal{Y}$ in Figure 5.9.1, then this notion of projection agrees with the concept of orthogonal projection that was discussed on p. 322. The phrase "oblique projection" is sometimes used to emphasize the fact that $\mathcal{X}$ and $\mathcal{Y}$ are not orthogonal subspaces. In this text the word "projection" is synonymous with the term "oblique projection." If it is known that $\mathcal{X} \perp \mathcal{Y}$, then we explicitly say "orthogonal projection." Orthogonal projections are discussed in detail on p. 429.

Given a pair of complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\Re^n$ and an arbitrary vector $\mathbf{v} \in \Re^n = \mathcal{X} \oplus \mathcal{Y}$, how can the projection of $\mathbf{v}$ onto $\mathcal{X}$ be computed? One way is to build a ***projector*** (a projection operator) that is a matrix $\mathbf{P}_{n\times n}$ with the property that for each $\mathbf{v} \in \Re^n$, the product $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$. Let $\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-r}\}$ be respective bases for $\mathcal{X}$ and $\mathcal{Y}$ so that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^n$—recall (5.9.4). This guarantees that if the $\mathbf{x}_i$'s and $\mathbf{y}_i$'s are placed as columns in

$$\mathbf{B}_{n\times n} = \big[\mathbf{x}_1\,\mathbf{x}_2 \cdots \mathbf{x}_r \,|\, \mathbf{y}_1\,\mathbf{y}_2 \cdots \mathbf{y}_{n-r}\big] = \big[\mathbf{X}_{n\times r} \,|\, \mathbf{Y}_{n\times(n-r)}\big],$$

then $\mathbf{B}$ is nonsingular. If $\mathbf{P}_{n\times n}$ is to have the property that $\mathbf{Pv}$ is the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ for every $\mathbf{v} \in \Re^n$, then (5.9.3) implies that $\mathbf{Px}_i = \mathbf{x}_i$, $i = 1, 2, \ldots, r$ and $\mathbf{Py}_j = \mathbf{0}$, $j = 1, 2, \ldots, n-r$, so

$$\mathbf{PB} = \mathbf{P}\big[\mathbf{X} \,|\, \mathbf{Y}\big] = \big[\mathbf{PX} \,|\, \mathbf{PY}\big] = \big[\mathbf{X} \,|\, \mathbf{0}\big]$$

and, consequently,

$$\mathbf{P} = \big[\mathbf{X} \,|\, \mathbf{0}\big]\mathbf{B}^{-1} = \mathbf{B}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{B}^{-1}. \tag{5.9.5}$$

To argue that $\mathbf{Pv}$ is indeed the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$, set $\mathbf{x} = \mathbf{Pv}$ and $\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{v}$ and observe that $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where

$$\mathbf{x} = \mathbf{Pv} = \big[\mathbf{X} \,|\, \mathbf{0}\big]\mathbf{B}^{-1}\mathbf{v} \in R(\mathbf{X}) = \mathcal{X} \tag{5.9.6}$$

and

$$\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{B}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{B}^{-1}\mathbf{v} = \big[\mathbf{0} \,|\, \mathbf{Y}\big]\mathbf{B}^{-1}\mathbf{v} \in R(\mathbf{Y}) = \mathcal{Y}. \tag{5.9.7}$$

Is it possible that there can be more than one projector onto $\mathcal{X}$ along $\mathcal{Y}$? No, $\mathbf{P}$ is unique because if $\mathbf{P}_1$ and $\mathbf{P}_2$ are two such projectors, then for $i = 1, 2$, we have $\mathbf{P}_i\mathbf{B} = \mathbf{P}_i\big[\mathbf{X}\,|\,\mathbf{Y}\big] = \big[\mathbf{P}_i\mathbf{X}\,|\,\mathbf{P}_i\mathbf{Y}\big] = \big[\mathbf{X}\,|\,\mathbf{0}\big]$, and this implies $\mathbf{P}_1\mathbf{B} = \mathbf{P}_2\mathbf{B}$, which means $\mathbf{P}_1 = \mathbf{P}_2$. Therefore, (5.9.5) is *the* projector onto $\mathcal{X}$ along $\mathcal{Y}$, and this formula for $\mathbf{P}$ is independent of which pair of bases for $\mathcal{X}$ and $\mathcal{Y}$ is selected. Notice that the argument involving (5.9.6) and (5.9.7) also establishes that the ***complementary projector***—the projector onto $\mathcal{Y}$ along $\mathcal{X}$—must be given by

$$\mathbf{Q} = \mathbf{I} - \mathbf{P} = \big[\mathbf{0}\,|\,\mathbf{Y}\big]\mathbf{B}^{-1} = \mathbf{B}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{B}^{-1}.$$

Below is a summary of the basic properties of projectors.

Projectors

Let $\mathcal{X}$ and $\mathcal{Y}$ be complementary subspaces of a vector space $\mathcal{V}$ so that each $\mathbf{v} \in \mathcal{V}$ can be uniquely resolved as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. The unique linear operator $\mathbf{P}$ defined by $\mathbf{P}\mathbf{v} = \mathbf{x}$ is called the ***projector onto*** $\mathcal{X}$ ***along*** $\mathcal{Y}$, and $\mathbf{P}$ has the following properties.

- $\mathbf{P}^2 = \mathbf{P}$ ($\mathbf{P}$ is idempotent). (5.9.8)
- $\mathbf{I} - \mathbf{P}$ is the complementary projector onto $\mathcal{Y}$ along $\mathcal{X}$. (5.9.9)
- $R(\mathbf{P}) = \{\mathbf{x}\,|\,\mathbf{P}\mathbf{x} = \mathbf{x}\}$ (the set of "fixed points" for $\mathbf{P}$). (5.9.10)
- $R(\mathbf{P}) = N(\mathbf{I} - \mathbf{P}) = \mathcal{X}$ and $R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P}) = \mathcal{Y}$. (5.9.11)
- If $\mathcal{V} = \Re^n$ or $\mathcal{C}^n$, then $\mathbf{P}$ is given by

$$\mathbf{P} = \big[\mathbf{X}\,|\,\mathbf{0}\big]\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1} = \big[\mathbf{X}\,|\,\mathbf{Y}\big]\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1}, \tag{5.9.12}$$

where the columns of $\mathbf{X}$ and $\mathbf{Y}$ are respective bases for $\mathcal{X}$ and $\mathcal{Y}$. Other formulas for $\mathbf{P}$ are given on p. 634.

Proof. Some of these properties have already been derived in the context of $\Re^n$. But since the concepts of projections and projectors are valid for all vector spaces, more general arguments that do not rely on properties of $\Re^n$ will be provided. Uniqueness is evident because if $\mathbf{P}_1$ and $\mathbf{P}_2$ both satisfy the defining condition, then $\mathbf{P}_1\mathbf{v} = \mathbf{P}_2\mathbf{v}$ for every $\mathbf{v} \in \mathcal{V}$, and thus $\mathbf{P}_1 = \mathbf{P}_2$. The linearity of $\mathbf{P}$ follows because if $\mathbf{v}_1 = \mathbf{x}_1 + \mathbf{y}_1$ and $\mathbf{v}_2 = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, then $\mathbf{P}(\alpha\mathbf{v}_1 + \mathbf{v}_2) = \alpha\mathbf{x}_1 + \mathbf{x}_2 = \alpha\mathbf{P}\mathbf{v}_1 + \mathbf{P}\mathbf{v}_2$. To prove that $\mathbf{P}$ is idempotent, write

$$\mathbf{P}^2\mathbf{v} = \mathbf{P}(\mathbf{P}\mathbf{v}) = \mathbf{P}\mathbf{x} = \mathbf{x} = \mathbf{P}\mathbf{v} \text{ for every } \mathbf{v} \in \mathcal{V} \implies \mathbf{P}^2 = \mathbf{P}.$$

The validity of (5.9.9) is established by observing that $\mathbf{v} = \mathbf{x} + \mathbf{y} = \mathbf{P}\mathbf{v} + \mathbf{y}$ implies $\mathbf{y} = \mathbf{v} - \mathbf{P}\mathbf{v} = (\mathbf{I} - \mathbf{P})\mathbf{v}$. The properties in (5.9.11) and (5.9.10) are immediate consequences of the definition. Formula (5.9.12) is the result of the arguments that culminated in (5.9.5), but it can be more elegantly derived by making use of the material in §4.7 and §4.8. If $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively, then $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n-r}\}$ is a basis for $\mathcal{V}$, and (4.7.4) says that the matrix of $\mathbf{P}$ with respect to $\mathcal{B}$ is

$$\begin{aligned}
[\mathbf{P}]_{\mathcal{B}} &= \Big[[\mathbf{P}\mathbf{x}_1]_{\mathcal{B}} \;\Big|\; \cdots \;\Big|\; [\mathbf{P}\mathbf{x}_r]_{\mathcal{B}} \;\Big|\; [\mathbf{P}\mathbf{y}_1]_{\mathcal{B}} \;\Big|\; \cdots \;\Big|\; [\mathbf{P}\mathbf{y}_{n-r}]_{\mathcal{B}} \Big] \\
&= \Big[[\mathbf{x}_1]_{\mathcal{B}} \;\Big|\; \cdots \;\Big|\; [\mathbf{x}_r]_{\mathcal{B}} \;\Big|\; [\mathbf{0}]_{\mathcal{B}} \;\Big|\; \cdots \;\Big|\; [\mathbf{0}]_{\mathcal{B}} \Big] \\
&= \Big[\mathbf{e}_1 \;\Big|\; \cdots \;\Big|\; \mathbf{e}_r \;\Big|\; \mathbf{0} \;\Big|\; \cdots \;\Big|\; \mathbf{0} \Big] = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.
\end{aligned}$$

If $\mathcal{S}$ is the standard basis, then (4.8.5) says that $[\mathbf{P}]_{\mathcal{B}} = \mathbf{B}^{-1}[\mathbf{P}]_{\mathcal{S}}\mathbf{B}$ in which

$$\mathbf{B} = [\mathbf{I}]_{\mathcal{B}\mathcal{S}} = \Big[[\mathbf{x}_1]_{\mathcal{S}} \;\Big|\; \cdots \;\Big|\; [\mathbf{x}_r]_{\mathcal{S}} \;\Big|\; [\mathbf{y}_1]_{\mathcal{S}} \;\cdots\; [\mathbf{y}_{n-r}]_{\mathcal{S}} \Big] = \big[\mathbf{X} \,|\, \mathbf{Y}\big],$$

and therefore $[\mathbf{P}]_{\mathcal{S}} = \mathbf{B}[\mathbf{P}]_{\mathcal{B}}\mathbf{B}^{-1} = \big[\mathbf{X} \,|\, \mathbf{Y}\big] \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \big[\mathbf{X} \,|\, \mathbf{Y}\big]^{-1}$. ∎

In the language of §4.8, statement (5.9.12) says that $\mathbf{P}$ is *similar* to the diagonal matrix $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. In the language of §4.9, this means that $\mathbf{P}$ must be the matrix representation of the linear operator that when restricted to $\mathcal{X}$ is the identity operator and when restricted to $\mathcal{Y}$ is the zero operator.

Statement (5.9.8) says that if $\mathbf{P}$ is a projector, then $\mathbf{P}$ is idempotent ($\mathbf{P}^2 = \mathbf{P}$). But what about the converse—is every idempotent linear operator necessarily a projector? The following theorem says, "Yes."

Projectors and Idempotents

A linear operator $\mathbf{P}$ on $\mathcal{V}$ is a projector if and only if $\mathbf{P}^2 = \mathbf{P}$. (5.9.13)

Proof. The fact that every projector is idempotent was proven in (5.9.8). The proof of the converse rests on the fact that

$$\mathbf{P}^2 = \mathbf{P} \implies R(\mathbf{P}) \text{ and } N(\mathbf{P}) \text{ are complementary subspaces.} \tag{5.9.14}$$

To prove this, observe that $\mathcal{V} = R(\mathbf{P}) + N(\mathbf{P})$ because for each $\mathbf{v} \in \mathcal{V}$,

$$\mathbf{v} = \mathbf{P}\mathbf{v} + (\mathbf{I} - \mathbf{P})\mathbf{v}, \quad \text{where} \quad \mathbf{P}\mathbf{v} \in R(\mathbf{P}) \text{ and } (\mathbf{I} - \mathbf{P})\mathbf{v} \in N(\mathbf{P}). \tag{5.9.15}$$

Furthermore, $R(\mathbf{P}) \cap N(\mathbf{P}) = \mathbf{0}$ because

$$\mathbf{x} \in R(\mathbf{P}) \cap N(\mathbf{P}) \implies \mathbf{x} = \mathbf{P}\mathbf{y} \text{ and } \mathbf{P}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{P}\mathbf{y} = \mathbf{P}^2\mathbf{y} = \mathbf{0},$$

and thus (5.9.14) is established. Now that we know $R(\mathbf{P})$ and $N(\mathbf{P})$ are complementary, we can conclude that $\mathbf{P}$ is a projector because each $\mathbf{v} \in \mathcal{V}$ can be uniquely written as $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in R(\mathbf{P})$ and $\mathbf{y} \in N(\mathbf{P})$, and (5.9.15) guarantees $\mathbf{P}\mathbf{v} = \mathbf{x}$. ∎

Notice that there is a one-to-one correspondence between the set of idempotents (or projectors) defined on a vector space $\mathcal{V}$ and the set of all pairs of complementary subspaces of $\mathcal{V}$ in the following sense.

- Each idempotent $\mathbf{P}$ defines a pair of complementary spaces—namely, $R(\mathbf{P})$ and $N(\mathbf{P})$.
- Every pair of complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ defines an idempotent—namely, the projector onto $\mathcal{X}$ along $\mathcal{Y}$.

Example 5.9.1

Problem: Let $\mathcal{X}$ and $\mathcal{Y}$ be the subspaces of $\Re^3$ that are spanned by

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \right\},$$

respectively. Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary, and then determine the projector onto $\mathcal{X}$ along $\mathcal{Y}$. What is the projection of $\mathbf{v} = \begin{pmatrix} -2 & 1 & 3 \end{pmatrix}^T$ onto $\mathcal{X}$ along $\mathcal{Y}$? What is the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$?

Solution: $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are linearly independent, so they are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are complementary because

$$rank\,[\mathbf{X} \,|\, \mathbf{Y}] = rank \left(\begin{array}{cc|c} 1 & 0 & 1 \\ -1 & 1 & -1 \\ -1 & -2 & 0 \end{array} \right) = 3$$

insures that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^3$—recall (5.9.4). The projector onto $\mathcal{X}$ along $\mathcal{Y}$ is obtained from (5.9.12) as

$$\mathbf{P} = \big[\mathbf{X} \,|\, \mathbf{0}\big]\big[\mathbf{X} \,|\, \mathbf{Y}\big]^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -2 & 0 \end{pmatrix} \begin{pmatrix} -2 & -2 & -1 \\ 1 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & -2 & -1 \\ 3 & 3 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

You may wish to verify that $\mathbf{P}$ is indeed idempotent. The projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ is $\mathbf{P}\mathbf{v}$, and, according to (5.9.9), the projection of $\mathbf{v}$ onto $\mathcal{Y}$ along $\mathcal{X}$ is $(\mathbf{I} - \mathbf{P})\mathbf{v}$.

Example 5.9.2

Angle between Complementary Subspaces. The angle between nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ in $\Re^n$ was defined on p. 295 to be the number $0 \leq \theta \leq \pi/2$ such that $\cos\theta = \mathbf{v}^T\mathbf{u}/\|\mathbf{v}\|_2\|\mathbf{u}\|_2$. It's natural to try to extend this idea to somehow make sense of angles between subspaces of $\Re^n$. Angles between completely general subspaces are presently out of our reach—they are discussed in §5.15—but the angle between a pair of *complementary* subspaces is within our grasp. When $\Re^n = \mathcal{R} \oplus \mathcal{N}$ with $\mathcal{R} \neq \mathbf{0} \neq \mathcal{N}$, the ***angle*** (also known as the ***minimal angle***) between $\mathcal{R}$ and $\mathcal{N}$ is defined to be the number $0 < \theta \leq \pi/2$ that satisfies

$$\cos\theta = \max_{\substack{\mathbf{u}\in\mathcal{R}\\ \mathbf{v}\in\mathcal{N}}} \frac{\mathbf{v}^T\mathbf{u}}{\|\mathbf{v}\|_2\|\mathbf{u}\|_2} = \max_{\substack{\mathbf{u}\in\mathcal{R},\,\mathbf{v}\in\mathcal{N}\\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u}. \tag{5.9.16}$$

While this is a good definition, it's not easy to use—especially if one wants to compute the numerical value of $\cos\theta$. The trick in making θ more accessible is to think in terms of projections and $\sin\theta = (1-\cos^2\theta)^{1/2}$. Let $\mathbf{P}$ be the projector such that $R(\mathbf{P}) = \mathcal{R}$ and $N(\mathbf{P}) = \mathcal{N}$, and recall that the matrix 2-norm (p. 281) of $\mathbf{P}$ is

$$\|\mathbf{P}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{P}\mathbf{x}\|_2. \tag{5.9.17}$$

In other words, $\|\mathbf{P}\|_2$ is the length of a longest vector in the image of the unit sphere under transformation by $\mathbf{P}$. To understand how $\sin\theta$ is related to $\|\mathbf{P}\|_2$, consider the situation in $\Re^3$. The image of the unit sphere under $\mathbf{P}$ is obtained by projecting the sphere onto $\mathcal{R}$ along lines parallel to $\mathcal{N}$. As depicted in Figure 5.9.2, the result is an ellipse in $\mathcal{R}$.

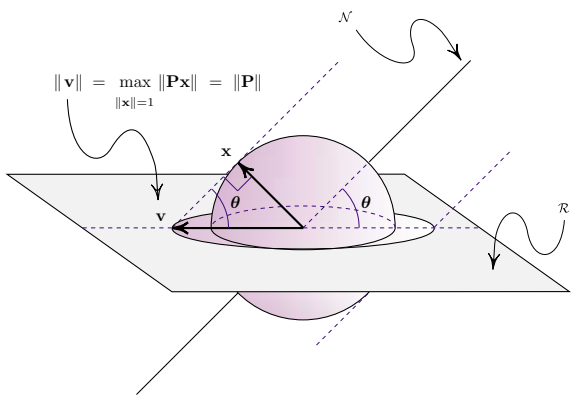


FIGURE 5.9.2

The norm of a longest vector $\mathbf{v}$ on this ellipse equals the norm of $\mathbf{P}$. That is, $\|\mathbf{v}\|_2 = \max_{\|\mathbf{x}\|_2=1}\|\mathbf{P}\mathbf{x}\|_2 = \|\mathbf{P}\|_2$, and it is apparent from the right triangle in

Figure 5.9.2 that

$$\sin\theta = \frac{\|\mathbf{x}\|_2}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{v}\|_2} = \frac{1}{\|\mathbf{P}\|_2}. \tag{5.9.18}$$

A little reflection on the geometry associated with Figure 5.9.2 should convince you that in $\Re^3$ a number θ satisfies (5.9.16) if and only if θ satisfies (5.9.18)—a completely rigorous proof validating this fact in $\Re^n$ is given in §5.15.

Note: Recall from p. 281 that $\|\mathbf{P}\|_2 = \sqrt{\lambda_{max}}$, where λ_{max} is the largest number λ such that $\mathbf{P}^T\mathbf{P} - \lambda\mathbf{I}$ is a singular matrix. Consequently,

$$\sin\theta = \frac{1}{\|\mathbf{P}\|_2} = \frac{1}{\sqrt{\lambda_{max}}}.$$

Numbers λ such that $\mathbf{P}^T\mathbf{P} - \lambda\mathbf{I}$ is singular are called *eigenvalues* of $\mathbf{P}^T\mathbf{P}$ (they are the main topic of discussion in Chapter 7, p. 489), and the numbers $\sqrt{\lambda}$ are the *singular values* of $\mathbf{P}$ discussed on p. 411.

Exercises for section 5.9

5.9.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of $\Re^3$ whose respective bases are

$$\mathcal{B}_{\mathcal{X}} = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \right\} \quad \text{and} \quad \mathcal{B}_{\mathcal{Y}} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\}.$$

(a) Explain why $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces of $\Re^3$.
(b) Determine the projector $\mathbf{P}$ onto $\mathcal{X}$ along $\mathcal{Y}$ as well as the complementary projector $\mathbf{Q}$ onto $\mathcal{Y}$ along $\mathcal{X}$.
(c) Determine the projection of $\mathbf{v} = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$ onto $\mathcal{Y}$ along $\mathcal{X}$.
(d) Verify that $\mathbf{P}$ and $\mathbf{Q}$ are both idempotent.
(e) Verify that $R(\mathbf{P}) = \mathcal{X} = N(\mathbf{Q})$ and $N(\mathbf{P}) = \mathcal{Y} = R(\mathbf{Q})$.

5.9.2. Construct an example of a pair of nontrivial complementary subspaces of $\Re^5$, and explain why your example is valid.

5.9.3. Construct an example to show that if $\mathcal{V} = \mathcal{X} + \mathcal{Y}$ but $\mathcal{X} \cap \mathcal{Y} \neq \mathbf{0}$, then a vector $\mathbf{v} \in \mathcal{V}$ can have two different representations as

$$\mathbf{v} = \mathbf{x}_1 + \mathbf{y}_1 \quad \text{and} \quad \mathbf{v} = \mathbf{x}_2 + \mathbf{y}_2,$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, but $\mathbf{x}_1 \neq \mathbf{x}_2$ and $\mathbf{y}_1 \neq \mathbf{y}_2$.

5.9.4. Explain why $\Re^{n\times n} = \mathcal{S} \oplus \mathcal{K}$, where $\mathcal{S}$ and $\mathcal{K}$ are the subspaces of $n \times n$ symmetric and skew-symmetric matrices, respectively. What is the projection of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ onto $\mathcal{S}$ along $\mathcal{K}$? **Hint:** Recall Exercise 3.2.6.

5.9.5. For a general vector space, let $\mathcal{X}$ and $\mathcal{Y}$ be two subspaces with respective bases $\mathcal{B}_{\mathcal{X}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$.

(a) Prove that $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$ if and only if $\{\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{y}_1, \ldots, \mathbf{y}_n\}$ is a linearly independent set.

(b) Does $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ being linear independent imply $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$?

(c) If $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a linearly independent set, does it follow that $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces? Why?

5.9.6. Let $\mathbf{P}$ be a projector defined on a vector space $\mathcal{V}$. Prove that (5.9.10) is true—i.e., prove that the range of a projector is the set of its "fixed points" in the sense that $R(\mathbf{P}) = \{\mathbf{x} \in \mathcal{V} \,|\, \mathbf{P}\mathbf{x} = \mathbf{x}\}$.

5.9.7. Suppose that $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, and let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$. Prove that (5.9.11) is true—i.e., prove

$$R(\mathbf{P}) = N(\mathbf{I} - \mathbf{P}) = \mathcal{X} \quad \text{and} \quad R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P}) = \mathcal{Y}.$$

5.9.8. Explain why $\|\mathbf{P}\|_2 \geq 1$ for every projector $\mathbf{P} \neq \mathbf{0}$. When is $\|\mathbf{P}\|_2 = 1$?

5.9.9. Explain why $\|\mathbf{I} - \mathbf{P}\|_2 = \|\mathbf{P}\|_2$ for all projectors that are not zero and not equal to the identity.

5.9.10. Prove that if $\mathbf{u}, \mathbf{v} \in \Re^{n\times 1}$ are vectors such that $\mathbf{v}^T\mathbf{u} = 1$, then

$$\left\|\mathbf{I} - \mathbf{u}\mathbf{v}^T\right\|_2 = \left\|\mathbf{u}\mathbf{v}^T\right\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 = \left\|\mathbf{u}\mathbf{v}^T\right\|_F,$$

where $\|\star\|_F$ is the Frobenius matrix norm defined in (5.2.1) on p. 279.

5.9.11. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are complementary subspaces of $\Re^n$, and let $\mathbf{B} = [\mathbf{X} \,|\, \mathbf{Y}]$ be a nonsingular matrix in which the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute respective bases for $\mathcal{X}$ and $\mathcal{Y}$. For an arbitrary vector $\mathbf{v} \in \Re^{n\times 1}$, explain why the projection of $\mathbf{v}$ onto $\mathcal{X}$ along $\mathcal{Y}$ can be obtained by the following two-step process.

(1) Solve the system $\mathbf{B}\mathbf{z} = \mathbf{v}$ for $\mathbf{z}$.

(2) Partition $\mathbf{z}$ as $\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{pmatrix}$, and set $\mathbf{p} = \mathbf{X}\mathbf{z}_1$.

5.9.12. Let $\mathbf{P}$ and $\mathbf{Q}$ be projectors.

(a) Prove $R(\mathbf{P}) = R(\mathbf{Q})$ if and only if $\mathbf{PQ} = \mathbf{Q}$ and $\mathbf{QP} = \mathbf{P}$.

(b) Prove $N(\mathbf{P}) = N(\mathbf{Q})$ if and only if $\mathbf{PQ} = \mathbf{P}$ and $\mathbf{QP} = \mathbf{Q}$.

(c) Prove that if $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_k$ are projectors with the same range, and if $\alpha_1, \alpha_2, \ldots, \alpha_k$ are scalars such that $\sum_j \alpha_j = 1$, then $\sum_j \alpha_j \mathbf{E}_j$ is a projector.

5.9.13. Prove that $rank(\mathbf{P}) = trace(\mathbf{P})$ for every projector $\mathbf{P}$ defined on $\Re^n$. **Hint:** Recall Example 3.6.5 (p. 110).

5.9.14. Let $\{\mathcal{X}_i\}_{i=1}^k$ be a collection of subspaces from a vector space $\mathcal{V}$, and let $\mathcal{B}_i$ denote a basis for $\mathcal{X}_i$. Prove that the following statements are equivalent.

(i) $\mathcal{V} = \mathcal{X}_1 + \mathcal{X}_2 + \cdots + \mathcal{X}_k$ and $\mathcal{X}_j \cap (\mathcal{X}_1 + \cdots + \mathcal{X}_{j-1}) = \mathbf{0}$ for each $j = 2, 3, \ldots, k$.

(ii) For each vector $\mathbf{v} \in \mathcal{V}$, there is one and only one way to write $\mathbf{v} = \mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k$, where $\mathbf{x}_i \in \mathcal{X}_i$.

(iii) $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ with $\mathcal{B}_i \cap \mathcal{B}_j = \phi$ for $i \neq j$ is a basis for $\mathcal{V}$.

Whenever any one of the above statements is true, $\mathcal{V}$ is said to be the ***direct sum*** of the $\mathcal{X}_i$'s, and we write $\mathcal{V} = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \cdots \oplus \mathcal{X}_k$. Notice that for $k = 2$, (i) and (5.9.1) say the same thing, and (ii) and (iii) reduce to (5.9.3) and (5.9.4), respectively.

5.9.15. For complementary subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\Re^n$, let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$, and let $\mathbf{Q} = [\mathbf{X} \,|\, \mathbf{Y}]$ in which the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. Prove that if $\mathbf{Q}^{-1}\mathbf{A}_{n\times n}\mathbf{Q}$ is partitioned as $\mathbf{Q}^{-1}\mathbf{AQ} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$, then

$$\mathbf{Q}\begin{pmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{PAP}, \qquad \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{PA}(\mathbf{I} - \mathbf{P}),$$

$$\mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = (\mathbf{I} - \mathbf{P})\mathbf{AP}, \quad \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{12} \end{pmatrix}\mathbf{Q}^{-1} = (\mathbf{I} - \mathbf{P})\mathbf{A}(\mathbf{I} - \mathbf{P}).$$

This means that if $\mathbf{A}$ is considered as a linear operator on $\Re^n$, and if $\mathcal{B} = \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$, where $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{Y}}$ are the respective bases for $\mathcal{X}$ and $\mathcal{Y}$ defined by the columns of $\mathbf{X}$ and $\mathbf{Y}$, then, in the context of §4.8, the matrix representation of $\mathbf{A}$ with respect to $\mathcal{B}$ is $[\mathbf{A}]_{\mathcal{B}} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$

in which the blocks are matrix representations of restricted operators as shown below.

$$\mathbf{A}_{11} = \left[\mathbf{PAP}_{/\mathcal{X}}\right]_{\mathcal{B}_\mathcal{X}}. \qquad \mathbf{A}_{12} = \left[\mathbf{PA}(\mathbf{I}-\mathbf{P})_{/\mathcal{Y}}\right]_{\mathcal{B}_\mathcal{Y}\mathcal{B}_\mathcal{X}}.$$

$$\mathbf{A}_{21} = \left[(\mathbf{I}-\mathbf{P})\mathbf{AP}_{/\mathcal{X}}\right]_{\mathcal{B}_\mathcal{X}\mathcal{B}_\mathcal{Y}}. \quad \mathbf{A}_{22} = \left[(\mathbf{I}-\mathbf{P})\mathbf{A}(\mathbf{I}-\mathbf{P})_{/\mathcal{Y}}\right]_{\mathcal{B}_\mathcal{Y}}.$$

5.9.16. Suppose that $\Re^n = \mathcal{X} \oplus \mathcal{Y}$, where $\dim \mathcal{X} = r$, and let $\mathbf{P}$ be the projector onto $\mathcal{X}$ along $\mathcal{Y}$. Explain why there exist matrices $\mathbf{X}_{n\times r}$ and $\mathbf{A}_{r\times n}$ such that $\mathbf{P} = \mathbf{XA}$, where $rank\,(\mathbf{X}) = rank\,(\mathbf{A}) = r$ and $\mathbf{AX} = \mathbf{I}_r$. This is a *full-rank factorization* for $\mathbf{P}$ (recall Exercise 3.9.8).

5.9.17. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that $\mathbf{E}+\mathbf{F}$ is a projector if and only if $\mathbf{EF} = \mathbf{FE} = \mathbf{0}$, and under this condition, prove that $R\,(\mathbf{E}+\mathbf{F}) = \mathcal{X}_1 \oplus \mathcal{X}_2$ and $N\,(\mathbf{E}+\mathbf{F}) = \mathcal{Y}_1 \cap \mathcal{Y}_2$.

5.9.18. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that $\mathbf{E}-\mathbf{F}$ is a projector if and only if $\mathbf{EF} = \mathbf{FE} = \mathbf{F}$, and under this condition, prove that $R\,(\mathbf{E}-\mathbf{F}) = \mathcal{X}_1 \cap \mathcal{Y}_2$ and $N\,(\mathbf{E}-\mathbf{F}) = \mathcal{Y}_1 \oplus \mathcal{X}_2$. **Hint:** $\mathbf{P}$ is a projector if and only if $\mathbf{I}-\mathbf{P}$ is a projector.

5.9.19. For either a real or complex vector space, let $\mathbf{E}$ be the projector onto $\mathcal{X}_1$ along $\mathcal{Y}_1$, and let $\mathbf{F}$ be the projector onto $\mathcal{X}_2$ along $\mathcal{Y}_2$. Prove that if $\mathbf{EF} = \mathbf{P} = \mathbf{FE}$, then $\mathbf{P}$ is the projector onto $\mathcal{X}_1 \cap \mathcal{X}_2$ along $\mathcal{Y}_1 + \mathcal{Y}_2$.

5.9.20. An ***inner pseudoinverse*** for $\mathbf{A}_{m\times n}$ is a matrix $\mathbf{X}_{n\times m}$ such that $\mathbf{AXA} = \mathbf{A}$, and an ***outer pseudoinverse*** for $\mathbf{A}$ is a matrix $\mathbf{X}$ satisfying $\mathbf{XAX} = \mathbf{X}$. When $\mathbf{X}$ is both an inner and outer pseudoinverse, $\mathbf{X}$ is called a ***reflexive pseudoinverse.***

(a) If $\mathbf{Ax} = \mathbf{b}$ is a consistent system of m equations in n unknowns, and if $\mathbf{A}^-$ is any inner pseudoinverse for $\mathbf{A}$, explain why the set of all solutions to $\mathbf{Ax} = \mathbf{b}$ can be expressed as

$$\mathbf{A}^-\mathbf{b} + R\left(\mathbf{I} - \mathbf{A}^-\mathbf{A}\right) = \{\mathbf{A}^-\mathbf{b} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{h} \,|\, \mathbf{h} \in \Re^n\}.$$

(b) Let $\mathcal{M}$ and $\mathcal{L}$ be respective complements of $R\,(\mathbf{A})$ and $N\,(\mathbf{A})$ so that $\mathcal{C}^m = R\,(\mathbf{A}) \oplus \mathcal{M}$ and $\mathcal{C}^n = \mathcal{L} \oplus N\,(\mathbf{A})$. Prove that there is a unique reflexive pseudoinverse $\mathbf{X}$ for $\mathbf{A}$ such that $R\,(\mathbf{X}) = \mathcal{L}$ and $N\,(\mathbf{X}) = \mathcal{M}$. Show that $\mathbf{X} = \mathbf{QA}^-\mathbf{P}$, where $\mathbf{A}^-$ is any inner pseudoinverse for $\mathbf{A}$, $\mathbf{P}$ is the projector onto $R\,(\mathbf{A})$ along $\mathcal{M}$, and $\mathbf{Q}$ is the projector onto $\mathcal{L}$ along $N\,(\mathbf{A})$.

5.10 RANGE-NULLSPACE DECOMPOSITION

Since there are infinitely many different pairs of complementary subspaces in $\Re^n$ (or $\mathcal{C}^n$),[54] is some pair more "natural" than the rest? Without reference to anything else the question is hard to answer. But if we start with a given matrix $\mathbf{A}_{n\times n}$, then there is a very natural direct sum decomposition of $\Re^n$ defined by fundamental subspaces associated with powers of $\mathbf{A}$. The rank plus nullity theorem on p. 199 says that $\dim R(\mathbf{A}) + \dim N(\mathbf{A}) = n$, so it's reasonable to ask about the possibility of $R(\mathbf{A})$ and $N(\mathbf{A})$ being complementary subspaces. If $\mathbf{A}$ is nonsingular, then it's trivially true that $R(\mathbf{A})$ and $N(\mathbf{A})$ are complementary, but when $\mathbf{A}$ is singular, this need not be the case because $R(\mathbf{A})$ and $N(\mathbf{A})$ need not be disjoint. For example,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \implies \begin{pmatrix} 1 \\ 0 \end{pmatrix} \in R(\mathbf{A}) \cap N(\mathbf{A}).$$

But all is not lost if we are willing to consider powers of $\mathbf{A}$.

Range-Nullspace Decomposition

For every singular matrix $\mathbf{A}_{n\times n}$, there exists a positive integer k such that $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are complementary subspaces. That is,

$$\Re^n = R\left(\mathbf{A}^k\right) \oplus N\left(\mathbf{A}^k\right). \tag{5.10.1}$$

The smallest positive integer k for which (5.10.1) holds is called the ***index*** of $\mathbf{A}$. For nonsingular matrices we define $index(\mathbf{A}) = 0$.

Proof. First observe that as $\mathbf{A}$ is powered the nullspaces grow and the ranges shrink—recall Exercise 4.2.12.

$$\begin{aligned} &N\left(\mathbf{A}^0\right) \subseteq N(\mathbf{A}) \subseteq N\left(\mathbf{A}^2\right) \subseteq \cdots \subseteq N\left(\mathbf{A}^k\right) \subseteq N\left(\mathbf{A}^{k+1}\right) \subseteq \cdots \\ &R\left(\mathbf{A}^0\right) \supseteq R(\mathbf{A}) \supseteq R\left(\mathbf{A}^2\right) \supseteq \cdots \supseteq R\left(\mathbf{A}^k\right) \supseteq R\left(\mathbf{A}^{k+1}\right) \supseteq \cdots. \end{aligned} \tag{5.10.2}$$

The proof of (5.10.1) is attained by combining the four following properties.

Property 1. *There is equality at some point in each of the chains* (5.10.2).

Proof. If there is strict containment at each link in the nullspace chain in (5.10.2), then the sequence of inequalities

$$\dim N\left(\mathbf{A}^0\right) < \dim N(\mathbf{A}) < \dim N\left(\mathbf{A}^2\right) < \dim N\left(\mathbf{A}^3\right) < \cdots$$

[54] All statements and arguments in this section are phrased in terms of $\Re^n$, but everything we say has a trivial extension to $\mathcal{C}^n$.

holds, and this forces $n < \dim N\left(\mathbf{A}^{n+1}\right)$, which is impossible. A similar argument proves equality exists somewhere in the range chain.

Property 2. *Once equality is attained, it is maintained throughout the rest of both chains in* (5.10.2). *In other words,*

$$\begin{aligned} &N\left(\mathbf{A}^0\right) \subset N(\mathbf{A}) \subset \cdots \subset N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right) = N\left(\mathbf{A}^{k+2}\right) = \cdots \\ &R\left(\mathbf{A}^0\right) \supset R(\mathbf{A}) \supset \cdots \supset R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^{k+2}\right) = \cdots. \end{aligned} \tag{5.10.3}$$

To prove this for the range chain, observe that if k is the smallest nonnegative integer such that $R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right)$, then for all $i \geq 1$,

$$R\left(\mathbf{A}^{i+k}\right) = R\left(\mathbf{A}^i\mathbf{A}^k\right) = \mathbf{A}^i R\left(\mathbf{A}^k\right) = \mathbf{A}^i R\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^{i+k+1}\right).$$

The nullspace chain stops growing at exactly the same place the ranges stop shrinking because the rank plus nullity theorem (p. 199) insures that $\dim N\left(\mathbf{A}^p\right) = n - \dim R\left(\mathbf{A}^p\right)$.

Property 3. *If k is the value at which the ranges stop shrinking and the nullspaces stop growing in* (5.10.3), *then* $R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) = \mathbf{0}$.

Proof. If $\mathbf{x} \in R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right)$, then $\mathbf{A}^k\mathbf{y} = \mathbf{x}$ for some $\mathbf{y} \in \Re^n$, and $\mathbf{A}^k\mathbf{x} = \mathbf{0}$. Hence $\mathbf{A}^{2k}\mathbf{y} = \mathbf{A}^k\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{y} \in N\left(\mathbf{A}^{2k}\right) = N\left(\mathbf{A}^k\right) \Rightarrow \mathbf{x} = \mathbf{0}$.

Property 4. *If k is the value at which the ranges stop shrinking and the nullspaces stop growing in* (5.10.3), *then* $R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right) = \Re^n$.

Proof. Use Property 3 along with (4.4.19), (4.4.15), and (4.4.6), to write

$$\begin{aligned} \dim\left[R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right)\right] &= \dim R\left(\mathbf{A}^k\right) + \dim N\left(\mathbf{A}^k\right) - \dim R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) \\ &= \dim R\left(\mathbf{A}^k\right) + \dim N\left(\mathbf{A}^k\right) = n \\ &\Longrightarrow \quad R\left(\mathbf{A}^k\right) + N\left(\mathbf{A}^k\right) = \Re^n. \quad \blacksquare \end{aligned}$$

Below is a summary of our observations concerning the index of a square matrix.

Index

The index of a square matrix $\mathbf{A}$ is the smallest nonnegative integer k such that any one of the three following statements is true.

- $rank\left(\mathbf{A}^k\right) = rank\left(\mathbf{A}^{k+1}\right)$.
- $R\left(\mathbf{A}^k\right) = R\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $R\left(\mathbf{A}^k\right)$ stops shrinking.
- $N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right)$—i.e., the point where $N\left(\mathbf{A}^k\right)$ stops growing.

For nonsingular matrices, $index(\mathbf{A}) = 0$. For singular matrices, $index(\mathbf{A})$ is the smallest *positive* integer k such that either of the following two statements is true.

- $R\left(\mathbf{A}^k\right) \cap N\left(\mathbf{A}^k\right) = \mathbf{0}$. (5.10.4)
- $\Re^n = R\left(\mathbf{A}^k\right) \oplus N\left(\mathbf{A}^k\right)$.

Example 5.10.1

Problem: Determine the index of $\mathbf{A} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{pmatrix}$.

Solution: $\mathbf{A}$ is singular (because $rank\,(\mathbf{A}) = 2$), so $index(\mathbf{A}) > 0$. Since

$$\mathbf{A}^2 = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^3 = \begin{pmatrix} 8 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

we see that $rank\,(\mathbf{A}) > rank\,(\mathbf{A}^2) = rank\,(\mathbf{A}^3)$, so $index(\mathbf{A}) = 2$. Alternately,

$$R\,(\mathbf{A}) = span\left\{ \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\},\ R\,(\mathbf{A}^2) = span \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix},\ R\,(\mathbf{A}^3) = span \begin{pmatrix} 8 \\ 0 \\ 0 \end{pmatrix},$$

so $R\,(\mathbf{A}) \supset R\,(\mathbf{A}^2) = R\,(\mathbf{A}^3)$ implies $index(\mathbf{A}) = 2$.

Nilpotent Matrices

- $\mathbf{N}_{n\times n}$ is said to be ***nilpotent*** whenever $\mathbf{N}^k = \mathbf{0}$ for some positive integer k.
- $k = index(\mathbf{N})$ is the smallest positive integer such that $\mathbf{N}^k = \mathbf{0}$. (Some authors refer to $index(\mathbf{N})$ as the ***index of nilpotency***.)

Proof. To prove that $k = index(\mathbf{N})$ is the smallest positive integer such that $\mathbf{N}^k = \mathbf{0}$, suppose p is a positive integer such that $\mathbf{N}^p = \mathbf{0}$, but $\mathbf{N}^{p-1} \neq \mathbf{0}$. We know from (5.10.3) that $R\left(\mathbf{N}^0\right) \supset R\,(\mathbf{N}) \supset \cdots \supset R\left(\mathbf{N}^k\right) = R\left(\mathbf{N}^{k+1}\right) = R\left(\mathbf{N}^{k+2}\right) = \cdots$, and this makes it clear that it's impossible to have $p < k$ or $p > k$, so $p = k$ is the only choice. ∎

Example 5.10.2

Problem: Verify that

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

is a nilpotent matrix, and determine its index.

Solution: Computing the powers

$$\mathbf{N}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{N}^3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

reveals that $\mathbf{N}$ is indeed nilpotent, and it shows that $index(\mathbf{N}) = 3$ because $\mathbf{N}^3 = \mathbf{0}$, but $\mathbf{N}^2 \neq \mathbf{0}$.

Anytime $\Re^n$ can be written as the direct sum of two complementary subspaces such that one of them is an invariant subspace for a given square matrix $\mathbf{A}$ we have a block-triangular representation for $\mathbf{A}$ according to formula (4.9.9) on p. 263. And if both complementary spaces are invariant under $\mathbf{A}$, then (4.9.10) says that this block-triangular representation is actually block diagonal.

Herein lies the true value of the range-nullspace decomposition (5.10.1) because it turns out that if $k = index(\mathbf{A})$, then $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are both invariant subspaces under $\mathbf{A}$. $R\left(\mathbf{A}^k\right)$ is invariant under $\mathbf{A}$ because

$$\mathbf{A}\big(R\left(\mathbf{A}^k\right)\big) = R\left(\mathbf{A}^{k+1}\right) = R\left(\mathbf{A}^k\right),$$

and $N\left(\mathbf{A}^k\right)$ is invariant because

$$\begin{aligned}
\mathbf{x} \in \mathbf{A}\big(N\left(\mathbf{A}^k\right)\big) \quad &\Longrightarrow \quad \mathbf{x} = \mathbf{A}\mathbf{w} \text{ for some } \mathbf{w} \in N\left(\mathbf{A}^k\right) = N\left(\mathbf{A}^{k+1}\right) \\
&\Longrightarrow \quad \mathbf{A}^k\mathbf{x} = \mathbf{A}^{k+1}\mathbf{w} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{x} \in N\left(\mathbf{A}^k\right) \\
&\Longrightarrow \quad \mathbf{A}\big(N\left(\mathbf{A}^k\right)\big) \subseteq N\left(\mathbf{A}^k\right).
\end{aligned}$$

This brings us to a matrix decomposition that is an important building block for developments that culminate in the Jordan form on p. 590.

Core-Nilpotent Decomposition

If $\mathbf{A}$ is an $n \times n$ singular matrix of index k such that $rank\left(\mathbf{A}^k\right) = r$, then there exists a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix} \tag{5.10.5}$$

in which $\mathbf{C}$ is nonsingular, and $\mathbf{N}$ is nilpotent of index k. In other words, $\mathbf{A}$ is *similar* to a 2×2 block-diagonal matrix containing a nonsingular "core" and a nilpotent component. The block-diagonal matrix in (5.10.5) is called a ***core-nilpotent decomposition*** of $\mathbf{A}$.

Note: When $\mathbf{A}$ is nonsingular, $k = 0$ and $r = n$, so $\mathbf{N}$ is not present, and we can set $\mathbf{Q} = \mathbf{I}$ and $\mathbf{C} = \mathbf{A}$ (the nonsingular core is everything). So (5.10.5) says absolutely nothing about nonsingular matrices.

Proof. Let $\mathbf{Q} = \big(\mathbf{X} \,|\, \mathbf{Y}\big)$, where the columns of $\mathbf{X}_{n\times r}$ and $\mathbf{Y}_{n\times n-r}$ constitute bases for $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$, respectively. Equation (4.9.10) guarantees that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ must be block diagonal in form, and thus (5.10.5) is established. To see that $\mathbf{N}$ is nilpotent, let

$$\mathbf{Q}^{-1} = \left(\begin{array}{c} \mathbf{U} \\ \hline \mathbf{V} \end{array}\right),$$

and write

$$\begin{pmatrix} \mathbf{C}^k & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^k \end{pmatrix} = \mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q} = \begin{pmatrix} \mathbf{U} \\ \hline \mathbf{V} \end{pmatrix} \mathbf{A}^k \left(\mathbf{X} \,|\, \mathbf{Y}\right) = \begin{pmatrix} \mathbf{U}\mathbf{A}^k\mathbf{X} & \mathbf{0} \\ \mathbf{V}\mathbf{A}^k\mathbf{X} & \mathbf{0} \end{pmatrix}.$$

Therefore, $\mathbf{N}^k = \mathbf{0}$ and $\mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q} = \begin{pmatrix} \mathbf{C}^k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Since $\mathbf{C}^k$ is $r \times r$ and $r = rank\left(\mathbf{A}^k\right) = rank\left(\mathbf{Q}^{-1}\mathbf{A}^k\mathbf{Q}\right) = rank\left(\mathbf{C}^k\right)$, it must be the case that $\mathbf{C}^k$ is nonsingular, and hence $\mathbf{C}$ is nonsingular. Finally, notice that $index(\mathbf{N}) = k$ because if $index(\mathbf{N}) \neq k$, then $\mathbf{N}^{k-1} = \mathbf{0}$, so

$$\begin{aligned} rank\left(\mathbf{A}^{k-1}\right) &= rank\left(\mathbf{Q}^{-1}\mathbf{A}^{k-1}\mathbf{Q}\right) = rank\begin{pmatrix} \mathbf{C}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^{k-1} \end{pmatrix} = rank\begin{pmatrix} \mathbf{C}^{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\ &= rank\left(\mathbf{C}^{k-1}\right) = r = rank\left(\mathbf{A}^k\right), \end{aligned}$$

which is impossible because $index(\mathbf{A}) = k$ is the smallest integer for which there is equality in ranks of powers. ∎

Example 5.10.3

Problem: Let $\mathbf{A}_{n\times n}$ have index k with $rank\left(\mathbf{A}^k\right) = r$, and let

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix} \quad \text{with} \quad \mathbf{Q} = \left(\mathbf{X}_{n\times r} \,|\, \mathbf{Y}\right) \text{ and } \mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{U}_{r\times n} \\ \hline \mathbf{V} \end{pmatrix}$$

be the core-nilpotent decomposition described in (5.10.5). Explain why

$$\mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{X}\mathbf{U} = \text{the projector onto } R\left(\mathbf{A}^k\right) \text{ along } N\left(\mathbf{A}^k\right)$$

and

$$\mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{Y}\mathbf{V} = \text{the projector onto } N\left(\mathbf{A}^k\right) \text{ along } R\left(\mathbf{A}^k\right).$$

Solution: Because $R\left(\mathbf{A}^k\right)$ and $N\left(\mathbf{A}^k\right)$ are complementary subspaces, and because the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute respective bases for these spaces, it follows from the discussion concerning projectors on p. 386 that

$$\mathbf{P} = \left(\mathbf{X} \,|\, \mathbf{Y}\right)\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\left(\mathbf{X} \,|\, \mathbf{Y}\right)^{-1} = \mathbf{Q}\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{X}\mathbf{U}$$

must be the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$, and

$$\mathbf{I} - \mathbf{P} = \left(\mathbf{X} \,|\, \mathbf{Y}\right)\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\left(\mathbf{X} \,|\, \mathbf{Y}\right)^{-1} = \mathbf{Q}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{pmatrix}\mathbf{Q}^{-1} = \mathbf{Y}\mathbf{V}$$

is the complementary projector onto $N\left(\mathbf{A}^k\right)$ along $R\left(\mathbf{A}^k\right)$.

Example 5.10.4

Problem: Explain how each noninvertible linear operator defined on an n-dimensional vector space $\mathcal{V}$ can be decomposed as the "direct sum" of an invertible operator and a nilpotent operator.

Solution: Let $\mathbf{T}$ be a linear operator of index k defined on $\mathcal{V} = \mathcal{R} \oplus \mathcal{N}$, where $\mathcal{R} = R\left(\mathbf{T}^k\right)$ and $\mathcal{N} = N\left(\mathbf{T}^k\right)$, and let $\mathbf{E} = \mathbf{T}_{/\mathcal{R}}$ and $\mathbf{F} = \mathbf{T}_{/\mathcal{N}}$ be the restriction operators as described in §4.9. Since $\mathcal{R}$ and $\mathcal{N}$ are invariant subspaces for $\mathbf{T}$, we know from the discussion of matrix representations on p. 263 that the right-hand side of the core-nilpotent decomposition in (5.10.5) must be the matrix representation of $\mathbf{T}$ with respect to a basis $\mathcal{B}_{\mathcal{R}} \cup \mathcal{B}_{\mathcal{N}}$, where $\mathcal{B}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{N}}$ are respective bases for $\mathcal{R}$ and $\mathcal{N}$. Furthermore, the nonsingular matrix $\mathbf{C}$ and the nilpotent matrix $\mathbf{N}$ are the matrix representations of $\mathbf{E}$ and $\mathbf{F}$ with respect to $\mathcal{B}_{\mathcal{R}}$ and $\mathcal{B}_{\mathcal{N}}$, respectively. Consequently, $\mathbf{E}$ is an invertible operator on $\mathcal{R}$, and $\mathbf{F}$ is a nilpotent operator on $\mathcal{N}$. Since $\mathcal{V} = \mathcal{R} \oplus \mathcal{N}$, each $\mathbf{x} \in \mathcal{V}$ can be expressed as $\mathbf{x} = \mathbf{r} + \mathbf{n}$ with $\mathbf{r} \in \mathcal{R}$ and $\mathbf{n} \in \mathcal{N}$. This allows us to formulate the concept of the ***direct sum*** of $\mathbf{E}$ and $\mathbf{F}$ by defining $\mathbf{E} \oplus \mathbf{F}$ to be the linear operator on $\mathcal{V}$ such that $(\mathbf{E} \oplus \mathbf{F})(\mathbf{x}) = \mathbf{E}(\mathbf{r}) + \mathbf{F}(\mathbf{n})$ for each $\mathbf{x} \in \mathcal{V}$. Therefore,

$$\begin{aligned}\mathbf{T}(\mathbf{x}) &= \mathbf{T}(\mathbf{r}+\mathbf{n}) = \mathbf{T}(\mathbf{r}) + \mathbf{T}(\mathbf{n}) = (\mathbf{T}_{/\mathcal{R}})(\mathbf{r}) + (\mathbf{T}_{/\mathcal{N}})(\mathbf{n}) \\ &= \mathbf{E}(\mathbf{r}) + \mathbf{F}(\mathbf{n}) = (\mathbf{E} \oplus \mathbf{F})(\mathbf{x}) \quad \text{for each} \quad \mathbf{x} \in \mathcal{V}.\end{aligned}$$

In other words, $\mathbf{T} = \mathbf{E} \oplus \mathbf{F}$ in which $\mathbf{E} = \mathbf{T}_{/\mathcal{R}}$ is invertible and $\mathbf{F} = \mathbf{T}_{/\mathcal{N}}$ is nilpotent.

Example 5.10.5

Drazin Inverse. Inverting the nonsingular core $\mathbf{C}$ and neglecting the nilpotent part $\mathbf{N}$ in the core-nilpotent decomposition (5.10.5) produces a natural generalization of matrix inversion. More precisely, if

$$\mathbf{A} = \mathbf{Q}\begin{pmatrix}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N}\end{pmatrix}\mathbf{Q}^{-1}, \quad \text{then} \quad \mathbf{A}^D = \mathbf{Q}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1} \tag{5.10.6}$$

defines the ***Drazin inverse*** of $\mathbf{A}$. Even though the components in a core-nilpotent decomposition are not uniquely defined by $\mathbf{A}$, it can be proven that $\mathbf{A}^D$ is unique and has the following properties.

- $\mathbf{A}^D = \mathbf{A}^{-1}$ when $\mathbf{A}$ is nonsingular (the nilpotent part is not present).
- $\mathbf{A}^D\mathbf{A}\mathbf{A}^D = \mathbf{A}^D$, $\mathbf{A}\mathbf{A}^D = \mathbf{A}^D\mathbf{A}$, $\mathbf{A}^{k+1}\mathbf{A}^D = \mathbf{A}^k$, where $k = index(\mathbf{A})$.[55]

[55] These three properties served as Michael P. Drazin's original definition in 1968. Initially,

- If $\mathbf{Ax} = \mathbf{b}$ is a consistent system of linear equations in which $\mathbf{b} \in R\left(\mathbf{A}^k\right)$, then $\mathbf{x} = \mathbf{A}^D\mathbf{b}$ is the unique solution that belongs to $R\left(\mathbf{A}^k\right)$ (Exercise 5.10.9).

- $\mathbf{AA}^D$ is the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$, and $\mathbf{I} - \mathbf{AA}^D$ is the complementary projector onto $N\left(\mathbf{A}^k\right)$ along $R\left(\mathbf{A}^k\right)$ (Exercise 5.10.10).

- If $\mathbf{A}$ is considered as a linear operator on $\Re^n$, then, with respect to a basis $\mathcal{B}_{\mathcal{R}}$ for $R\left(\mathbf{A}^k\right)$, $\mathbf{C}$ is the matrix representation for the restricted operator $\mathbf{A}/_{R(\mathbf{A}^k)}$ (see p. 263). Thus $\mathbf{A}/_{R(\mathbf{A}^k)}$ is invertible. Moreover,

$$\left[\mathbf{A}^D/_{R(\mathbf{A}^k)}\right]_{\mathcal{B}_{\mathcal{R}}} = \mathbf{C}^{-1} = \left[\left(\mathbf{A}/_{R(\mathbf{A}^k)}\right)^{-1}\right]_{\mathcal{B}_{\mathcal{R}}}, \quad \text{so} \quad \mathbf{A}^D/_{R(\mathbf{A}^k)} = \left(\mathbf{A}/_{R(\mathbf{A}^k)}\right)^{-1}.$$

  In other words, $\mathbf{A}^D$ is the inverse of $\mathbf{A}$ on $R\left(\mathbf{A}^k\right)$, and $\mathbf{A}^D$ is the zero operator on $N\left(\mathbf{A}^k\right)$, so, in the context of Example 5.10.4,

$$\mathbf{A} = \mathbf{A}/_{R(\mathbf{A}^k)} \oplus \mathbf{A}/_{N(\mathbf{A}^k)} \quad \text{and} \quad \mathbf{A}^D = \left(\mathbf{A}/_{R(\mathbf{A}^k)}\right)^{-1} \oplus \mathbf{0}/_{N(\mathbf{A}^k)}.$$

Exercises for section 5.10

5.10.1. If $\mathbf{A}$ is a square matrix of index $k > 0$, prove that $index(\mathbf{A}^k) = 1$.

5.10.2. If $\mathbf{A}$ is a nilpotent matrix of index k, describe the components in a core-nilpotent decomposition of $\mathbf{A}$.

5.10.3. Prove that if $\mathbf{A}$ is a symmetric matrix, then $index(\mathbf{A}) \leq 1$.

5.10.4. $\mathbf{A} \in \mathcal{C}^{n\times n}$ is said to be a ***normal matrix*** whenever $\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$. Prove that if $\mathbf{A}$ is normal, then $index(\mathbf{A}) \leq 1$.
Note: All symmetric matrices are normal, so the result of this exercise includes the result of Exercise 5.10.3 as a special case.

Drazin's concept attracted little interest—perhaps due to Drazin's abstract algebraic presentation. But eventually Drazin's generalized inverse was recognized to be a useful tool for analyzing nonorthogonal types of problems involving singular matrices. In this respect, the Drazin inverse is complementary to the Moore–Penrose pseudoinverse discussed in Exercise 4.5.20 and on p. 423 because the Moore–Penrose pseudoinverse is more useful in applications where orthogonality is somehow wired in (e.g., least squares).

5.10.5. Find a core-nilpotent decomposition and the Drazin inverse of

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

5.10.6. For a square matrix $\mathbf{A}$, any scalar λ that makes $\mathbf{A} - \lambda\mathbf{I}$ singular is called an ***eigenvalue*** for $\mathbf{A}$. The *index of an eigenvalue* λ is defined to be the index of the associated matrix $\mathbf{A} - \lambda\mathbf{I}$. In other words, $index(\lambda) = index(\mathbf{A} - \lambda\mathbf{I})$. Determine the eigenvalues and the index of each eigenvalue for the following matrices:

$$\text{(a)} \quad \mathbf{J} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}. \qquad \text{(b)} \quad \mathbf{J} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix}.$$

5.10.7. Let $\mathbf{P}$ be a projector different from the identity.
(a) Explain why $index(\mathbf{P}) = 1$. What is the index of $\mathbf{I}$?
(b) Determine the core-nilpotent decomposition for $\mathbf{P}$.

5.10.8. Let $\mathbf{N}$ be a nilpotent matrix of index k, and suppose that $\mathbf{x}$ is a vector such that $\mathbf{N}^{k-1}\mathbf{x} \neq \mathbf{0}$. Prove that the set

$$\mathcal{C} = \{\mathbf{x}, \mathbf{N}\mathbf{x}, \mathbf{N}^2\mathbf{x}, \ldots, \mathbf{N}^{k-1}\mathbf{x}\}$$

is a linearly independent set. $\mathcal{C}$ is sometimes called a ***Jordan chain*** or a ***Krylov sequence***.

5.10.9. Let $\mathbf{A}$ be a square matrix of index k, and let $\mathbf{b} \in R\left(\mathbf{A}^k\right)$.
(a) Explain why the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ must be consistent.
(b) Explain why $\mathbf{x} = \mathbf{A}^D\mathbf{b}$ is the unique solution in $R\left(\mathbf{A}^k\right)$.
(c) Explain why the general solution is given by $\mathbf{A}^D\mathbf{b} + N\left(\mathbf{A}\right)$.

5.10.10. Suppose that $\mathbf{A}$ is a square matrix of index k, and let $\mathbf{A}^D$ be the Drazin inverse of $\mathbf{A}$ as defined in Example 5.10.5. Explain why $\mathbf{A}\mathbf{A}^D$ is the projector onto $R\left(\mathbf{A}^k\right)$ along $N\left(\mathbf{A}^k\right)$. What does $\mathbf{I} - \mathbf{A}\mathbf{A}^D$ project onto and along?

5.10.11. An ***algebraic group*** is a set $\mathcal{G}$ together with an associative operation between its elements such that $\mathcal{G}$ is closed with respect to this operation; $\mathcal{G}$ possesses an identity element $\mathbf{E}$ (which can be proven to be unique); and every member $\mathbf{A} \in \mathcal{G}$ has an inverse $\mathbf{A}^{\#}$ (which can be proven to be unique). These are essentially the axioms (A1), (A2), (A4), and (A5) in the definition of a vector space given on p. 160. A ***matrix group*** is a set of square matrices that forms an algebraic group under ordinary matrix multiplication.

(a) Show that the set of $n \times n$ nonsingular matrices is a matrix group.

(b) Show that the set of $n \times n$ unitary matrices is a *subgroup* of the $n \times n$ nonsingular matrices.

(c) Show that the set $\mathcal{G} = \left\{ \begin{pmatrix} \alpha & \alpha \\ \alpha & \alpha \end{pmatrix} \,\middle|\, \alpha \neq 0 \right\}$ is a matrix group. In particular, what does the identity element $\mathbf{E} \in \mathcal{G}$ look like, and what does the inverse $\mathbf{A}^{\#}$ of $\mathbf{A} \in \mathcal{G}$ look like?

5.10.12. For singular matrices, prove that the following statements are equivalent.

(a) $\mathbf{A}$ is a group matrix (i.e., $\mathbf{A}$ belongs to a matrix group).

(b) $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$.

(c) $R(\mathbf{A})$ and $N(\mathbf{A})$ are complementary subspaces.

(d) $index(\mathbf{A}) = 1$.

(e) There are nonsingular matrices $\mathbf{Q}_{n\times n}$ and $\mathbf{C}_{r\times r}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \text{where} \quad r = rank\,(\mathbf{A}).$$

5.10.13. Let $\mathbf{A} \in \mathcal{G}$ for some matrix group $\mathcal{G}$.

(a) Show that the identity element $\mathbf{E} \in \mathcal{G}$ is the projector onto $R(\mathbf{A})$ along $N(\mathbf{A})$ by arguing that $\mathbf{E}$ must be of the form

$$\mathbf{E} = \mathbf{Q} \begin{pmatrix} \mathbf{I}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}^{-1}.$$

(b) Show that the ***group inverse*** of $\mathbf{A}$ (the inverse of $\mathbf{A}$ in $\mathcal{G}$) must be of the form

$$\mathbf{A}^{\#} = \mathbf{Q} \begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q}^{-1}.$$

5.11 ORTHOGONAL DECOMPOSITION

The orthogonal complement of a single vector $\mathbf{x}$ was defined on p. 322 to be the set of all vectors orthogonal to $\mathbf{x}$. Below is the natural extension of this idea.

Orthogonal Complement

For a subset $\mathcal{M}$ of an inner-product space $\mathcal{V}$, the ***orthogonal complement*** $\mathcal{M}^{\perp}$ (pronounced "$\mathcal{M}$ perp") of $\mathcal{M}$ is defined to be the set of all vectors in $\mathcal{V}$ that are orthogonal to every vector in $\mathcal{M}$. That is,

$$\mathcal{M}^{\perp} = \left\{ \mathbf{x} \in \mathcal{V} \;\middle|\; \langle \mathbf{m} | \mathbf{x} \rangle = 0 \text{ for all } \mathbf{m} \in \mathcal{M} \right\}.$$

For example, if $\mathcal{M} = \{\mathbf{x}\}$ is a single vector in $\Re^2$, then, as illustrated in Figure 5.11.1, $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to $\mathbf{x}$. If $\mathcal{M}$ is a plane through the origin in $\Re^3$, then $\mathcal{M}^{\perp}$ is the line through the origin that is perpendicular to the plane.

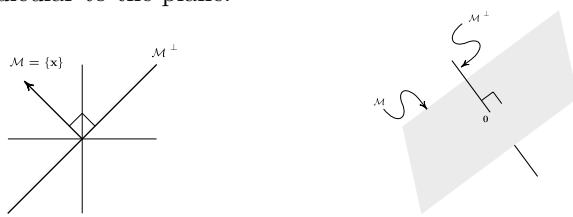


FIGURE 5.11.1

Notice that $\mathcal{M}^{\perp}$ is a subspace of $\mathcal{V}$ even if $\mathcal{M}$ is not a subspace because $\mathcal{M}^{\perp}$ is closed with respect to vector addition and scalar multiplication (Exercise 5.11.4). But if $\mathcal{M}$ is a subspace, then $\mathcal{M}$ and $\mathcal{M}^{\perp}$ decompose $\mathcal{V}$ as described below.

Orthogonal Complementary Subspaces

If $\mathcal{M}$ is a subspace of a finite-dimensional inner-product space $\mathcal{V}$, then

$$\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^{\perp}. \tag{5.11.1}$$

Furthermore, if $\mathcal{N}$ is a subspace such that $\mathcal{V} = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{N} \perp \mathcal{M}$ (every vector in $\mathcal{N}$ is orthogonal to every vector in $\mathcal{M}$), then

$$\mathcal{N} = \mathcal{M}^{\perp}. \tag{5.11.2}$$

Proof. Observe that $\mathcal{M} \cap \mathcal{M}^{\perp} = \mathbf{0}$ because if $\mathbf{x} \in \mathcal{M}$ and $\mathbf{x} \in \mathcal{M}^{\perp}$, then $\mathbf{x}$ must be orthogonal to itself, and $\langle \mathbf{x} | \mathbf{x} \rangle = 0$ implies $\mathbf{x} = 0$. To prove that $\mathcal{M} \oplus \mathcal{M}^{\perp} = \mathcal{V}$, suppose that $\mathcal{B}_{\mathcal{M}}$ and $\mathcal{B}_{\mathcal{M}^{\perp}}$ are orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively. Since $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are disjoint, $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}}$ is an orthonormal basis for some subspace $\mathcal{S} = \mathcal{M} \oplus \mathcal{M}^{\perp} \subseteq \mathcal{V}$. If $\mathcal{S} \neq \mathcal{V}$, then the basis extension technique of Example 4.4.5 followed by the Gram–Schmidt orthogonalization procedure of §5.5 yields a nonempty set of vectors $\mathcal{E}$ such that $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}} \cup \mathcal{E}$ is an orthonormal basis for $\mathcal{V}$. Consequently,

$$\mathcal{E} \perp \mathcal{B}_{\mathcal{M}} \implies \mathcal{E} \perp \mathcal{M} \implies \mathcal{E} \subseteq \mathcal{M}^{\perp} \implies \mathcal{E} \subseteq span\left(\mathcal{B}_{\mathcal{M}^{\perp}}\right).$$

But this is impossible because $\mathcal{B}_{\mathcal{M}} \cup \mathcal{B}_{\mathcal{M}^{\perp}} \cup \mathcal{E}$ is linearly independent. Therefore, $\mathcal{E}$ is the empty set, and thus $\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^{\perp}$. To prove statement (5.11.2), note that $\mathcal{N} \perp \mathcal{M}$ implies $\mathcal{N} \subseteq \mathcal{M}^{\perp}$, and coupling this with the fact that $\mathcal{M} \oplus \mathcal{M}^{\perp} = \mathcal{V} = \mathcal{M} \oplus \mathcal{N}$ together with (4.4.19) insures

$$\dim \mathcal{N} = \dim \mathcal{V} - \dim \mathcal{M} = \dim \mathcal{M}^{\perp}. \quad \blacksquare$$

Example 5.11.1

Problem: Let $\mathbf{U}_{m \times m} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$ be a partitioned orthogonal matrix. Explain why $R(\mathbf{U}_1)$ and $R(\mathbf{U}_2)$ must be orthogonal complements of each other.

Solution: Statement (5.9.4) insures that $\Re^m = R(\mathbf{U}_1) \oplus R(\mathbf{U}_2)$, and we know that $R(\mathbf{U}_1) \perp R(\mathbf{U}_2)$ because the columns of $\mathbf{U}$ are an orthonormal set. Therefore, (5.11.2) guarantees that $R(\mathbf{U}_2) = R(\mathbf{U}_1)^{\perp}$.

Perp Operation

If $\mathcal{M}$ is a subspace of an n-dimensional inner-product space, then the following statements are true.

- $\dim \mathcal{M}^{\perp} = n - \dim \mathcal{M}.$ (5.11.3)
- $\mathcal{M}^{\perp^{\perp}} = \mathcal{M}.$ (5.11.4)

Proof. Property (5.11.3) follows from the fact that $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are complementary subspaces—recall (4.4.19). To prove (5.11.4), first show that $\mathcal{M}^{\perp^{\perp}} \subseteq \mathcal{M}$. If $\mathbf{x} \in \mathcal{M}^{\perp^{\perp}}$, then (5.11.1) implies $\mathbf{x} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^{\perp}$, so

$$0 = \langle \mathbf{n} | \mathbf{x} \rangle = \langle \mathbf{n} | \mathbf{m} + \mathbf{n} \rangle = \langle \mathbf{n} | \mathbf{m} \rangle + \langle \mathbf{n} | \mathbf{n} \rangle = \langle \mathbf{n} | \mathbf{n} \rangle \implies \mathbf{n} = 0 \implies \mathbf{x} \in \mathcal{M},$$

and thus $\mathcal{M}^{\perp^{\perp}} \subseteq \mathcal{M}$. We know from (5.11.3) that $\dim \mathcal{M}^{\perp} = n - \dim \mathcal{M}$ and $\dim \mathcal{M}^{\perp^{\perp}} = n - \dim \mathcal{M}^{\perp}$, so $\dim \mathcal{M}^{\perp^{\perp}} = \dim \mathcal{M}$. Therefore, (4.4.6) guarantees that $\mathcal{M}^{\perp^{\perp}} = \mathcal{M}$. $\blacksquare$

We are now in a position to understand why the four fundamental subspaces associated with a matrix $\mathbf{A} \in \Re^{m\times n}$ are indeed "fundamental." First consider $R(\mathbf{A})^{\perp}$, and observe that for all $\mathbf{y} \in \Re^n$,

$$\begin{aligned}
\mathbf{x} \in R(\mathbf{A})^{\perp} &\iff \langle \mathbf{A}\mathbf{y} | \mathbf{x} \rangle = 0 \iff \mathbf{y}^T\mathbf{A}^T\mathbf{x} = 0 \\
&\iff \left\langle \mathbf{y} \middle| \mathbf{A}^T\mathbf{x} \right\rangle = 0 \iff \mathbf{A}^T\mathbf{x} = \mathbf{0} \quad \text{(Exercise 5.3.2)} \\
&\iff \mathbf{x} \in N\left(\mathbf{A}^T\right).
\end{aligned}$$

Therefore, $R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right)$. Perping both sides of this equation and replacing[56] $\mathbf{A}$ by $\mathbf{A}^T$ produces $R\left(\mathbf{A}^T\right) = N(\mathbf{A})^{\perp}$. Combining these observations produces one of the fundamental theorems of linear algebra.

Orthogonal Decomposition Theorem

For every $\mathbf{A} \in \Re^{m\times n}$,

$$R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right) \quad \text{and} \quad N(\mathbf{A})^{\perp} = R\left(\mathbf{A}^T\right). \tag{5.11.5}$$

In light of (5.11.1), this means that every matrix $\mathbf{A} \in \Re^{m\times n}$ produces an orthogonal decomposition of $\Re^m$ and $\Re^n$ in the sense that

$$\Re^m = R(\mathbf{A}) \oplus R(\mathbf{A})^{\perp} = R(\mathbf{A}) \oplus N\left(\mathbf{A}^T\right), \tag{5.11.6}$$

and

$$\Re^n = N(\mathbf{A}) \oplus N(\mathbf{A})^{\perp} = N(\mathbf{A}) \oplus R\left(\mathbf{A}^T\right). \tag{5.11.7}$$

Theorems without hypotheses tend to be extreme in the sense that they either say very little or they reveal a lot. The orthogonal decomposition theorem has no hypothesis—it holds for all matrices—so, does it really say something significant? Yes, it does, and here's part of the reason why.

In addition to telling us how to decompose $\Re^m$ and $\Re^n$ in terms of the four fundamental subspaces of $\mathbf{A}$, the orthogonal decomposition theorem also tells us how to decompose $\mathbf{A}$ itself into more basic components. Suppose that $rank(\mathbf{A}) = r$, and let

$$\mathcal{B}_{R(\mathbf{A})} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\} \quad \text{and} \quad \mathcal{B}_{N\left(\mathbf{A}^T\right)} = \{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$$

be orthonormal bases for $R(\mathbf{A})$ and $N\left(\mathbf{A}^T\right)$, respectively, and let

$$\mathcal{B}_{R\left(\mathbf{A}^T\right)} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\} \quad \text{and} \quad \mathcal{B}_{N(\mathbf{A})} = \{\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \ldots, \mathbf{v}_n\}$$

[56] Here, as well as throughout the rest of this section, $(\star)^T$ can be replaced by $(\star)^*$ whenever $\Re^{m\times n}$ is replaced by $\mathcal{C}^{m\times n}$.

be orthonormal bases for $R\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})$, respectively. It follows that $\mathcal{B}_{R(\mathbf{A})} \cup \mathcal{B}_{N\left(\mathbf{A}^T\right)}$ and $\mathcal{B}_{R\left(\mathbf{A}^T\right)} \cup \mathcal{B}_{N(\mathbf{A})}$ are orthonormal bases for $\Re^m$ and $\Re^n$, respectively, and hence

$$\mathbf{U}_{m\times m} = \Big(\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_m\Big) \quad \text{and} \quad \mathbf{V}_{n\times n} = \Big(\mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{v}_n\Big) \tag{5.11.8}$$

are orthogonal matrices. Now consider the product $\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V}$, and notice that $r_{ij} = \mathbf{u}_i^T\mathbf{A}\mathbf{v}_j$. However, $\mathbf{u}_i^T\mathbf{A} = \mathbf{0}$ for $i = r+1, \ldots, m$ and $\mathbf{A}\mathbf{v}_j = \mathbf{0}$ for $j = r+1, \ldots, n$, so

$$\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{u}_1^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_1^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ \mathbf{u}_r^T\mathbf{A}\mathbf{v}_1 & \cdots & \mathbf{u}_r^T\mathbf{A}\mathbf{v}_r & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{5.11.9}$$

In other words, $\mathbf{A}$ can be factored as

$$\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{V}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T. \tag{5.11.10}$$

Moreover, $\mathbf{C}$ is nonsingular because it is $r \times r$ and

$$rank\,(\mathbf{C}) = rank\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = rank\left(\mathbf{U}^T\mathbf{A}\mathbf{V}\right) = rank\,(\mathbf{A}) = r.$$

For lack of a better name, we will refer to (5.11.10) as a ***URV factorization.***

We have just observed that every set of orthonormal bases for the four fundamental subspaces defines a URV factorization. The situation is also reversible in the sense that every URV factorization of $\mathbf{A}$ defines an orthonormal basis for each fundamental subspace. Starting with orthogonal matrices $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ together with a nonsingular matrix $\mathbf{C}_{r\times r}$ such that (5.11.10) holds, use the fact that right-hand multiplication by a nonsingular matrix does not alter the range (Exercise 4.5.12) to observe

$$R\,(\mathbf{A}) = R\,(\mathbf{U}\mathbf{R}) = R\,(\mathbf{U}_1\mathbf{C} \,|\, \mathbf{0}) = R\,(\mathbf{U}_1\mathbf{C}) = R\,(\mathbf{U}_1).$$

By (5.11.5) and Example 5.11.1, $N\left(\mathbf{A}^T\right) = R\,(\mathbf{A})^\perp = R\,(\mathbf{U}_1)^\perp = R\,(\mathbf{U}_2)$. Similarly, left-hand multiplication by a nonsingular matrix does not change the nullspace, so the second equation in (5.11.5) along with Example 5.11.1 yields

$$N\,(\mathbf{A}) = N\left(\mathbf{R}\mathbf{V}^T\right) = N\begin{pmatrix} \mathbf{C}\mathbf{V}_1^T \\ \mathbf{0} \end{pmatrix} = N\left(\mathbf{C}\mathbf{V}_1^T\right) = N\left(\mathbf{V}_1^T\right) = R\,(\mathbf{V}_1)^\perp = R\,(\mathbf{V}_2),$$

and $R\left(\mathbf{A}^T\right) = N\,(\mathbf{A})^\perp = R\,(\mathbf{V}_2)^\perp = R\,(\mathbf{V}_1)$. A summary is given below.

URV Factorization

For each $\mathbf{A} \in \Re^{m\times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m\times m}$ and $\mathbf{V}_{n\times n}$ and a nonsingular matrix $\mathbf{C}_{r\times r}$ such that

$$\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{V}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T. \tag{5.11.11}$$

- The first r columns in $\mathbf{U}$ are an orthonormal basis for $R(\mathbf{A})$.
- The last $m-r$ columns of $\mathbf{U}$ are an orthonormal basis for $N(\mathbf{A}^T)$.
- The first r columns in $\mathbf{V}$ are an orthonormal basis for $R(\mathbf{A}^T)$.
- The last $n-r$ columns of $\mathbf{V}$ are an orthonormal basis for $N(\mathbf{A})$.

Each different collection of orthonormal bases for the four fundamental subspaces of $\mathbf{A}$ produces a different URV factorization of $\mathbf{A}$. In the complex case, replace $(\star)^T$ by $(\star)^*$ and "orthogonal" by "unitary."

Example 5.11.2

Problem: Explain how to make $\mathbf{C}$ lower triangular in (5.11.11).

Solution: Apply Householder (or Givens) reduction to produce an orthogonal matrix $\mathbf{P}_{m\times m}$ such that $\mathbf{PA} = \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{B}$ is $r \times n$ of rank r. Householder (or Givens) reduction applied to $\mathbf{B}^T$ results in an orthogonal matrix $\mathbf{Q}_{n\times n}$ and a nonsingular upper-triangular matrix $\mathbf{T}$ such that

$$\mathbf{Q}\mathbf{B}^T = \begin{pmatrix} \mathbf{T}_{r\times r} \\ \mathbf{0} \end{pmatrix} \implies \mathbf{B} = (\mathbf{T}^T \,|\, \mathbf{0})\mathbf{Q} \implies \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q},$$

so $\mathbf{A} = \mathbf{P}^T\begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix} = \mathbf{P}^T\begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}$ is a URV factorization.

Note: $\mathbf{C}$ can in fact be made diagonal—see (p. 412).

Have you noticed the duality that has emerged concerning the use of fundamental subspaces of $\mathbf{A}$ to decompose $\Re^n$ (or $\mathcal{C}^n$)? On one hand there is the range-nullspace decomposition (p. 394), and on the other is the orthogonal decomposition theorem (p. 405). Each produces a decomposition of $\mathbf{A}$. The range-nullspace decomposition of $\Re^n$ produces the core-nilpotent decomposition of $\mathbf{A}$ (p. 397), and the orthogonal decomposition theorem produces the URV factorization. In the next section, the URV factorization specializes to become

the singular value decomposition (p. 412), and in a somewhat parallel manner, the core-nilpotent decomposition paves the way to the Jordan form (p. 590). These two parallel tracks constitute the backbone for the theory of modern linear algebra, so it's worthwhile to take a moment and reflect on them.

The range-nullspace decomposition decomposes $\Re^n$ with *square* matrices while the orthogonal decomposition theorem does it with *rectangular* matrices. So does this mean that the range-nullspace decomposition is a special case of, or somehow weaker than, the orthogonal decomposition theorem? No! Even for square matrices they are not very comparable because each says something that the other doesn't. The core-nilpotent decomposition (and eventually the Jordan form) is obtained by a similarity transformation, and, as discussed in §§4.8–4.9, similarity is the primary mechanism for revealing characteristics of $\mathbf{A}$ that are independent of bases or coordinate systems. The URV factorization has little to say about such things because it is generally not a similarity transformation. Orthogonal decomposition has the advantage whenever orthogonality is naturally built into a problem—such as least squares applications. And, as discussed in §5.7, orthogonal methods often produce numerically stable algorithms for floating-point computation, whereas similarity transformations are generally not well suited for numerical computations. The value of similarity is mainly on the theoretical side of the coin.

So when do we get the best of both worlds—i.e., when is a URV factorization also a core-nilpotent decomposition? First, $\mathbf{A}$ must be square and, second, (5.11.11) must be a similarity transformation, so $\mathbf{U} = \mathbf{V}$. Surprisingly, this happens for a rather large class of matrices described below.

Range Perpendicular to Nullspace

For $rank\,(\mathbf{A}_{n\times n}) = r$, the following statements are equivalent:

- $R(\mathbf{A}) \perp N(\mathbf{A})$, (5.11.12)
- $R(\mathbf{A}) = R\left(\mathbf{A}^T\right)$, (5.11.13)
- $N(\mathbf{A}) = N\left(\mathbf{A}^T\right)$, (5.11.14)
- $\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T$ (5.11.15)

in which $\mathbf{U}$ is orthogonal and $\mathbf{C}$ is nonsingular. Such matrices will be called ***RPN matrices,*** short for "range perpendicular to nullspace." Some authors call them *range-symmetric* or *EP* matrices. Nonsingular matrices are trivially RPN because they have a zero nullspace. For complex matrices, replace $(\star)^T$ by $(\star)^*$ and "orthogonal" by "unitary."

Proof. The fact that (5.11.12) $\Longleftrightarrow$ (5.11.13) $\Longleftrightarrow$ (5.11.14) is a direct consequence of (5.11.5). It suffices to prove (5.11.15) $\Longleftrightarrow$ (5.11.13). If (5.11.15) is a

URV factorization with $\mathbf{V} = \mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$, then $R(\mathbf{A}) = R(\mathbf{U}_1) = R(\mathbf{V}_1) = R\left(\mathbf{A}^T\right)$. Conversely, if $R(\mathbf{A}) = R\left(\mathbf{A}^T\right)$, perping both sides and using equation (5.11.5) produces $N(\mathbf{A}) = N\left(\mathbf{A}^T\right)$, so (5.11.8) yields a URV factorization with $\mathbf{U} = \mathbf{V}$. ∎

Example 5.11.3

$\mathbf{A} \in \mathcal{C}^{n\times n}$ is called a ***normal matrix*** whenever $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$. As illustrated in Figure 5.11.2, normal matrices fill the niche between hermitian and (complex) RPN matrices in the sense that real-symmetric $\Rightarrow$ hermitian $\Rightarrow$ normal $\Rightarrow$ RPN, with no implication being reversible—details are called for in Exercise 5.11.13.

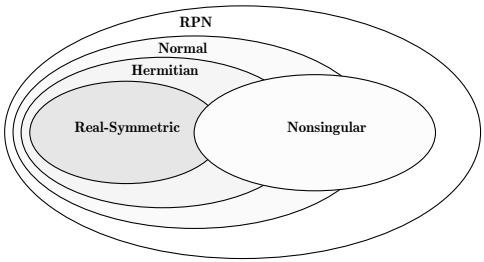


FIGURE 5.11.2

Exercises for section 5.11

5.11.1. Verify the orthogonal decomposition theorem for $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ -1 & -1 & 0 \\ -2 & -1 & -1 \end{pmatrix}$.

5.11.2. For an inner-product space $\mathcal{V}$, what is $\mathcal{V}^\perp$? What is $\mathbf{0}^\perp$?

5.11.3. Find a basis for the orthogonal complement of $\mathcal{M} = span\left\{\begin{pmatrix} 1 \\ 2 \\ 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 1 \\ 6 \end{pmatrix}\right\}$.

5.11.4. For every inner-product space $\mathcal{V}$, prove that if $\mathcal{M} \subseteq \mathcal{V}$, then $\mathcal{M}^\perp$ is a subspace of $\mathcal{V}$.

5.11.5. If $\mathcal{M}$ and $\mathcal{N}$ are subspaces of an n-dimensional inner-product space, prove that the following statements are true.

(a) $\mathcal{M} \subseteq \mathcal{N} \implies \mathcal{N}^\perp \subseteq \mathcal{M}^\perp$.
(b) $(\mathcal{M} + \mathcal{N})^\perp = \mathcal{M}^\perp \cap \mathcal{N}^\perp$.
(c) $(\mathcal{M} \cap \mathcal{N})^\perp = \mathcal{M}^\perp + \mathcal{N}^\perp$.

5.11.6. Explain why the rank plus nullity theorem on p. 199 is a corollary of the orthogonal decomposition theorem.

5.11.7. Suppose $\mathbf{A} = \mathbf{URV}^T$ is a URV factorization of an $m \times n$ matrix of rank r, and suppose $\mathbf{U}$ is partitioned as $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$, where $\mathbf{U}_1$ is $m \times r$. Prove that $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T$ is the projector onto $R(\mathbf{A})$ along $N\left(\mathbf{A}^T\right)$. In this case, $\mathbf{P}$ is said to be an *orthogonal projector* because its range is orthogonal to its nullspace. What is the orthogonal projector onto $N\left(\mathbf{A}^T\right)$ along $R(\mathbf{A})$? (Orthogonal projectors are discussed in more detail on p. 429.)

5.11.8. Use the Householder reduction method as described in Example 5.11.2 to compute a URV factorization as well as orthonormal bases for the four fundamental subspaces of $\mathbf{A} = \begin{pmatrix} -4 & -2 & -4 & -2 \\ 2 & -2 & 2 & 1 \\ -4 & 1 & -4 & -2 \end{pmatrix}$.

5.11.9. Compute a URV factorization for the matrix given in Exercise 5.11.8 by using elementary row operations together with Gram–Schmidt orthogonalization. Are the results the same as those of Exercise 5.11.8?

5.11.10. For the matrix $\mathbf{A}$ of Exercise 5.11.8, find vectors $\mathbf{x} \in R(\mathbf{A})$ and $\mathbf{y} \in N\left(\mathbf{A}^T\right)$ such that $\mathbf{v} = \mathbf{x} + \mathbf{y}$, where $\mathbf{v} = \begin{pmatrix} 3 & 3 & 3 \end{pmatrix}^T$. Is there more than one choice for $\mathbf{x}$ and $\mathbf{y}$?

5.11.11. Construct a square matrix such that $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$, but $R(\mathbf{A})$ is not orthogonal to $N(\mathbf{A})$.

5.11.12. For $\mathbf{A}_{n\times n}$ singular, explain why $R(\mathbf{A}) \perp N(\mathbf{A})$ implies $index(\mathbf{A}) = 1$, but not conversely.

5.11.13. Prove that real-symmetric matrix $\Rightarrow$ hermitian $\Rightarrow$ normal $\Rightarrow$ (complex) RPN. Construct examples to show that none of the implications is reversible.

5.11.14. Let $\mathbf{A}$ be a normal matrix.

(a) Prove that $R(\mathbf{A} - \lambda\mathbf{I}) \perp N(\mathbf{A} - \lambda\mathbf{I})$ for every scalar λ.

(b) Let λ and μ be scalars such that $\mathbf{A} - \lambda\mathbf{I}$ and $\mathbf{A} - \mu\mathbf{I}$ are singular matrices—such scalars are called ***eigenvalues*** of $\mathbf{A}$. Prove that if $\lambda \neq \mu$, then $N(\mathbf{A} - \lambda\mathbf{I}) \perp N(\mathbf{A} - \mu\mathbf{I})$.

5.12 SINGULAR VALUE DECOMPOSITION

For an $m \times n$ matrix $\mathbf{A}$ of rank r, Example 5.11.2 shows how to build a URV factorization

$$\mathbf{A} = \mathbf{URV}^T = \mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T$$

in which $\mathbf{C}$ is triangular. The purpose of this section is to prove that it's possible to do even better by showing that $\mathbf{C}$ can be made to be *diagonal*. To see how, let $\sigma_1 = \|\mathbf{A}\|_2 = \|\mathbf{C}\|_2$ (Exercise 5.6.9), and recall from the proof of (5.2.7) on p. 281 that $\|\mathbf{C}\|_2 = \|\mathbf{Cx}\|_2$ for some vector $\mathbf{x}$ such that

$$(\mathbf{C}^T\mathbf{C} - \lambda\mathbf{I})\mathbf{x} = 0, \quad \text{where} \quad \|\mathbf{x}\|_2 = 1 \text{ and } \lambda = \mathbf{x}^T\mathbf{C}^T\mathbf{Cx} = \sigma_1^2. \tag{5.12.1}$$

Set $\mathbf{y} = \mathbf{Cx}/\|\mathbf{Cx}\|_2 = \mathbf{Cx}/\sigma_1$, and let $\mathbf{R}_y = \big(\mathbf{y}\,|\,\mathbf{Y}\big)$ and $\mathbf{R}_x = \big(\mathbf{x}\,|\,\mathbf{X}\big)$ be elementary reflectors having $\mathbf{y}$ and $\mathbf{x}$ as their first columns, respectively—recall Example 5.6.3. Reflectors are orthogonal matrices, so $\mathbf{x}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{Y}^T\mathbf{y} = \mathbf{0}$, and these together with (5.12.1) yield

$$\mathbf{y}^T\mathbf{CX} = \frac{\mathbf{x}^T\mathbf{C}^T\mathbf{CX}}{\sigma_1} = \frac{\lambda\mathbf{x}^T\mathbf{X}}{\sigma_1} = \mathbf{0} \quad \text{and} \quad \mathbf{Y}^T\mathbf{Cx} = \sigma_1\mathbf{Y}^T\mathbf{y} = \mathbf{0}.$$

Coupling these facts with $\mathbf{y}^T\mathbf{Cx} = \mathbf{y}^T(\sigma_1\mathbf{y}) = \sigma_1$ and $\mathbf{R}_y = \mathbf{R}_y^T$ produces

$$\mathbf{R}_y\mathbf{CR}_x = \begin{pmatrix} \mathbf{y}^T \\ \hline \mathbf{Y}^T \end{pmatrix}\mathbf{C}\big(\mathbf{x}\,|\,\mathbf{X}\big) = \begin{pmatrix} \mathbf{y}^T\mathbf{Cx} & \mathbf{y}^T\mathbf{CX} \\ \mathbf{Y}^T\mathbf{Cx} & \mathbf{Y}^T\mathbf{CX} \end{pmatrix} = \begin{pmatrix} \sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 \end{pmatrix}$$

with $\sigma_1 \geq \|\mathbf{C}_2\|_2$ (because $\sigma_1 = \|\mathbf{C}\|_2 = \max\{\sigma_1, \|\mathbf{C}_2\|\}$ by (5.2.12)). Repeating the process on $\mathbf{C}_2$ yields reflectors $\mathbf{S}_y$, $\mathbf{S}_x$ such that

$$\mathbf{S}_y\mathbf{C}_2\mathbf{S}_x = \begin{pmatrix} \sigma_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_3 \end{pmatrix}, \quad \text{where } \sigma_2 \geq \|\mathbf{C}_3\|_2\,.$$

If $\mathbf{P}_2$ and $\mathbf{Q}_2$ are the orthogonal matrices

$$\mathbf{P}_2 = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{S}_y \end{pmatrix}\mathbf{R}_y, \quad \mathbf{Q}_2 = \mathbf{R}_x\begin{pmatrix} 1 & 0 \\ 0 & \mathbf{S}_x \end{pmatrix}, \text{ then } \mathbf{P}_2\mathbf{CQ}_2 = \begin{pmatrix} \sigma_1 & 0 & \mathbf{0} \\ 0 & \sigma_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 \end{pmatrix}$$

in which $\sigma_1 \geq \sigma_2 \geq \|\mathbf{C}_3\|_2$. Continuing for $r-1$ times produces orthogonal matrices $\mathbf{P}_{r-1}$ and $\mathbf{Q}_{r-1}$ such that $\mathbf{P}_{r-1}\mathbf{CQ}_{r-1} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) = \mathbf{D}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. If $\tilde{\mathbf{U}}^T$ and $\tilde{\mathbf{V}}$ are the orthogonal matrices

$$\tilde{\mathbf{U}}^T = \begin{pmatrix} \mathbf{P}_{r-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\mathbf{U}^T \text{ and } \tilde{\mathbf{V}} = \mathbf{V}\begin{pmatrix} \mathbf{Q}_{r-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \text{ then } \tilde{\mathbf{U}}^T\mathbf{A}\tilde{\mathbf{V}} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and thus the *singular value decomposition* (SVD) is derived.[57]

[57] The SVD has been independently discovered and rediscovered several times. Those credited with the early developments include Eugenio Beltrami (1835–1899) in 1873, M. E. Camille Jordan (1838–1922) in 1875, James J. Sylvester (1814–1897) in 1889, L. Autonne in 1913, and C. Eckart and G. Young in 1936.

Singular Value Decomposition

For each $\mathbf{A} \in \Re^{m\times n}$ of rank r, there are orthogonal matrices $\mathbf{U}_{m\times m}$, $\mathbf{V}_{n\times n}$ and a diagonal matrix $\mathbf{D}_{r\times r} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ such that

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T \quad \text{with} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0. \tag{5.12.2}$$

The σ_i's are called the nonzero ***singular values*** of $\mathbf{A}$. When $r < p = \min\{m, n\}$, $\mathbf{A}$ is said to have $p - r$ additional zero singular values. The factorization in (5.12.2) is called a ***singular value decomposition*** of $\mathbf{A}$, and the columns in $\mathbf{U}$ and $\mathbf{V}$ are called left-hand and right-hand ***singular vectors*** for $\mathbf{A}$, respectively.

While the constructive method used to derive the SVD can be used as an algorithm, more sophisticated techniques exist, and all good matrix computation packages contain numerically stable SVD implementations. However, the details of a practical SVD algorithm are too complicated to be discussed at this point.

The SVD is valid for complex matrices when $(\star)^T$ is replaced by $(\star)^*$, and it can be shown that the singular values are unique, but the singular vectors are not. In the language of Chapter 7, the σ_i^2's are the eigenvalues of $\mathbf{A}^T\mathbf{A}$, and the singular vectors are specialized sets of eigenvectors for $\mathbf{A}^T\mathbf{A}$—see the summary on p. 555. In fact, the practical algorithm for computing the SVD is an implementation of the QR iteration (p. 535) that is cleverly applied to $\mathbf{A}^T\mathbf{A}$ without ever explicitly computing $\mathbf{A}^T\mathbf{A}$.

Singular values reveal something about the geometry of linear transformations because the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ of a matrix $\mathbf{A}$ tell us how much distortion can occur under transformation by $\mathbf{A}$. They do so by giving us an explicit picture of how $\mathbf{A}$ distorts the unit sphere. To develop this, suppose that $\mathbf{A} \in \Re^{n\times n}$ is nonsingular (Exercise 5.12.5 treats the singular and rectangular case), and let $\mathcal{S}_2 = \{\mathbf{x} \mid \|\mathbf{x}\|_2 = 1\}$ be the unit 2-sphere in $\Re^n$. The nature of the image $\mathbf{A}(\mathcal{S}_2)$ is revealed by considering the singular value decompositions

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad \text{and} \quad \mathbf{A}^{-1} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T \quad \text{with} \quad \mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices. For each $\mathbf{y} \in \mathbf{A}(\mathcal{S}_2)$ there is an $\mathbf{x} \in \mathcal{S}_2$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$, so, with $\mathbf{w} = \mathbf{U}^T\mathbf{y}$,

$$\begin{aligned} 1 &= \|\mathbf{x}\|_2^2 = \left\|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\right\|_2^2 = \left\|\mathbf{A}^{-1}\mathbf{y}\right\|_2^2 = \left\|\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 = \left\|\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 \\ &= \left\|\mathbf{D}^{-1}\mathbf{w}\right\|_2^2 = \frac{w_1^2}{\sigma_1^2} + \frac{w_2^2}{\sigma_2^2} + \cdots + \frac{w_r^2}{\sigma_r^2}. \end{aligned} \tag{5.12.3}$$

This means that $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is an ellipsoid whose k^{th} semiaxis has length σ_k. Because orthogonal transformations are isometries (length preserving transformations), $\mathbf{U}^T$ can only affect the orientation of $\mathbf{A}(\mathcal{S}_2)$, so $\mathbf{A}(\mathcal{S}_2)$ is also an ellipsoid whose k^{th} semiaxis has length σ_k. Furthermore, (5.12.3) implies that the ellipsoid $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is in standard position—i.e., its axes are directed along the standard basis vectors $\mathbf{e}_k$. Since $\mathbf{U}$ maps $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ to $\mathbf{A}(\mathcal{S}_2)$, and since $\mathbf{U}\mathbf{e}_k = \mathbf{U}_{*k}$, it follows that the axes of $\mathbf{A}(\mathcal{S}_2)$ are directed along the left-hand singular vectors defined by the columns of $\mathbf{U}$. Therefore, the kth semiaxis of $\mathbf{A}(\mathcal{S}_2)$ is $\sigma_k\mathbf{U}_{*k}$. Finally, since $\mathbf{AV} = \mathbf{UD}$ implies $\mathbf{AV}_{*k} = \sigma_k\mathbf{U}_{*k}$, the right-hand singular vector $\mathbf{V}_{*k}$ is a point on $\mathcal{S}_2$ that is mapped to the kth semiaxis vector on the ellipsoid $\mathbf{A}(\mathcal{S}_2)$. The picture in $\Re^3$ looks like Figure 5.12.1.

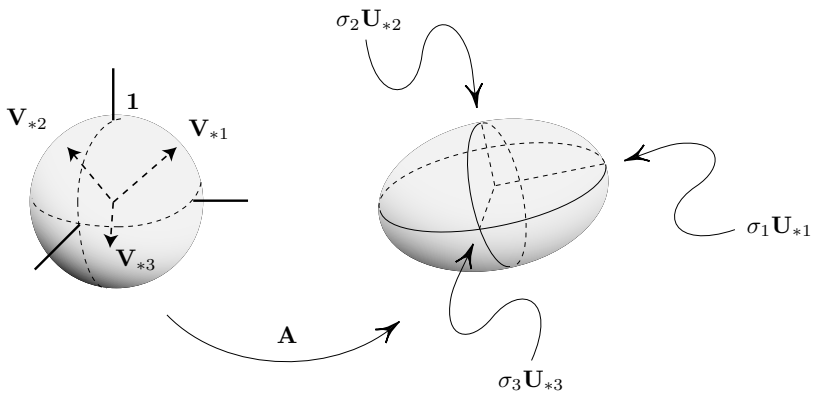


FIGURE 5.12.1

The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is therefore measured by $\kappa_2 = \sigma_1/\sigma_n$, the ratio of the largest singular value to the smallest singular value. Moreover, from the discussion of induced matrix norms (p. 280) and the unitary invariance of the 2-norm (Exercise 5.6.9),

$$\max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \|\mathbf{A}\|_2 = \left\|\mathbf{UDV}^T\right\|_2 = \|\mathbf{D}\|_2 = \sigma_1$$

and

$$\min_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \frac{1}{\|\mathbf{A}^{-1}\|_2} = \frac{1}{\|\mathbf{VD}^{-1}\mathbf{U}^T\|_2} = \frac{1}{\|\mathbf{D}^{-1}\|_2} = \sigma_n.$$

In other words, longest and shortest vectors on $\mathbf{A}(\mathcal{S}_2)$ have respective lengths $\sigma_1 = \|\mathbf{A}\|_2$ and $\sigma_n = 1/\left\|\mathbf{A}^{-1}\right\|_2$ (this justifies Figure 5.2.1 on p. 281), so $\kappa_2 = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2$. This is called the 2-*norm condition number* of $\mathbf{A}$. Different norms result in condition numbers with different values but with more or less the same order of magnitude as κ_2 (see Exercise 5.12.3), so the qualitative information about distortion is the same. Below is a summary.

Image of the Unit Sphere

For a nonsingular $\mathbf{A}_{n\times n}$ having singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ and an SVD $\mathbf{A} = \mathbf{UDV}^T$ with $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, the image of the unit 2-sphere is an ellipsoid whose k^{th} semiaxis is given by $\sigma_k \mathbf{U}_{*k}$ (see Figure 5.12.1). Furthermore, $\mathbf{V}_{*k}$ is a point on the unit sphere such that $\mathbf{AV}_{*k} = \sigma_k \mathbf{U}_{*k}$. In particular,

- $$\sigma_1 = \|\mathbf{AV}_{*1}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \|\mathbf{A}\|_2, \tag{5.12.4}$$
- $$\sigma_n = \|\mathbf{AV}_{*n}\|_2 = \min_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = 1/\|\mathbf{A}^{-1}\|_2. \tag{5.12.5}$$

The degree of distortion of the unit sphere under transformation by $\mathbf{A}$ is measured by the 2-norm ***condition number***

- $$\kappa_2 = \frac{\sigma_1}{\sigma_n} = \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2 \geq 1. \tag{5.12.6}$$

Notice that $\kappa_2 = 1$ if and only if $\mathbf{A}$ is an orthogonal matrix.

The amount of distortion of the unit sphere under transformation by $\mathbf{A}$ determines the degree to which uncertainties in a linear system $\mathbf{Ax} = \mathbf{b}$ can be magnified. This is explained in the following example.

Example 5.12.1

Uncertainties in Linear Systems. Systems of linear equations $\mathbf{Ax} = \mathbf{b}$ arising in practical work almost always come with built-in uncertainties due to modeling errors (because assumptions are almost always necessary), data collection errors (because infinitely precise gauges don't exist), and data entry errors (because numbers like $\sqrt{2}$, π, and $2/3$ can't be entered exactly). In addition, roundoff error in floating-point computation is a prevalent source of uncertainty. In all cases it's important to estimate the degree of uncertainty in the solution of $\mathbf{Ax} = \mathbf{b}$. This is not difficult when $\mathbf{A}$ is known exactly and all uncertainty resides in the right-hand side. Even if this is not the case, it's sometimes possible to aggregate uncertainties and shift all of them to the right-hand side.

Problem: Let $\mathbf{Ax} = \mathbf{b}$ be a nonsingular system in which $\mathbf{A}$ is known exactly but $\mathbf{b}$ is subject to an uncertainty $\mathbf{e}$, and consider $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{e} = \tilde{\mathbf{b}}$. Estimate the *relative uncertainty*[58] $\|\mathbf{x} - \tilde{\mathbf{x}}\| / \|\mathbf{x}\|$ in $\mathbf{x}$ in terms of the relative uncertainty $\|\mathbf{b} - \tilde{\mathbf{b}}\| / \|\mathbf{b}\| = \|\mathbf{e}\| / \|\mathbf{b}\|$ in $\mathbf{b}$. Use any vector norm and its induced matrix norm (p. 280).

[58] Knowing the *absolute* uncertainty $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ by itself may not be meaningful. For example, an absolute uncertainty of a half of an inch might be fine when measuring the distance between the earth and the moon, but it's not good in the practice of eye surgery.

Solution: Use $\|\mathbf{b}\| = \|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\,\|\mathbf{x}\|$ with $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{e}$ to write

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} = \frac{\left\|\mathbf{A}^{-1}\mathbf{e}\right\|}{\|\mathbf{x}\|} \le \frac{\|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|\|\mathbf{e}\|}{\|\mathbf{b}\|} = \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \tag{5.12.7}$$

where $\kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|$ is a *condition number* as discussed earlier ($\kappa = \sigma_1/\sigma_n$ if the 2-norm is used). Furthermore, $\|\mathbf{e}\| = \|\mathbf{A}(\mathbf{x}-\tilde{\mathbf{x}})\| \le \|\mathbf{A}\|\,\|(\mathbf{x}-\tilde{\mathbf{x}})\|$ and $\|\mathbf{x}\| \le \left\|\mathbf{A}^{-1}\right\|\|\mathbf{b}\|$ imply

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \ge \frac{\|\mathbf{e}\|}{\|\mathbf{A}\|\,\|\mathbf{x}\|} \ge \frac{\|\mathbf{e}\|}{\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\|\mathbf{b}\|} = \frac{1}{\kappa}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}.$$

This with (5.12.7) yields the following bounds on the relative uncertainty:

$$\kappa^{-1}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \le \frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^{-1}\right\|. \tag{5.12.8}$$

In other words, when $\mathbf{A}$ is *well conditioned* (i.e., when κ is small—see the rule of thumb in Example 3.8.2 to get a feeling of what "small" and "large" might mean), (5.12.8) insures that small relative uncertainties in $\mathbf{b}$ cannot greatly affect the solution, but when $\mathbf{A}$ is *ill conditioned* (i.e., when κ is large), a relatively small uncertainty in $\mathbf{b}$ *might* result in a relatively large uncertainty in $\mathbf{x}$. To be more sure, the following problem needs to be addressed.

Problem: Can equality be realized in each bound in (5.12.8) for every nonsingular $\mathbf{A}$, and if so, how?

Solution: Use the 2-norm, and let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be an SVD so $\mathbf{A}\mathbf{V}_{*k} = \sigma_k\mathbf{U}_{*k}$ for each k. If $\mathbf{b}$ and $\mathbf{e}$ are directed along left-hand singular vectors associated with σ_1 and σ_n, respectively—say, $\mathbf{b} = \beta\mathbf{U}_{*1}$ and $\mathbf{e} = \epsilon\mathbf{U}_{*n}$, then

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{A}^{-1}(\beta\mathbf{U}_{*1}) = \frac{\beta\mathbf{V}_{*1}}{\sigma_1} \quad \text{and} \quad \mathbf{x}-\tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{e} = \mathbf{A}^{-1}(\epsilon\mathbf{U}_{*n}) = \frac{\epsilon\mathbf{V}_{*n}}{\sigma_n},$$

so

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} = \left(\frac{\sigma_1}{\sigma_n}\right)\frac{|\epsilon|}{|\beta|} = \kappa_2\frac{\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \quad \text{when } \mathbf{b} = \beta\mathbf{U}_{*1} \text{ and } \mathbf{e} = \epsilon\mathbf{U}_{*n}.$$

Thus the upper bound (the worst case) in (5.12.8) is attainable for all $\mathbf{A}$. The lower bound (the best case) is realized in the opposite situation when $\mathbf{b}$ and $\mathbf{e}$ are directed along $\mathbf{U}_{*n}$ and $\mathbf{U}_{*1}$, respectively. If $\mathbf{b} = \beta\mathbf{U}_{*n}$ and $\mathbf{e} = \epsilon\mathbf{U}_{*1}$, then the same argument yields $\mathbf{x} = \sigma_n^{-1}\beta\mathbf{V}_{*n}$ and $\mathbf{x}-\tilde{\mathbf{x}} = \sigma_1^{-1}\epsilon\mathbf{V}_{*1}$, so

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2} = \left(\frac{\sigma_n}{\sigma_1}\right)\frac{|\epsilon|}{|\beta|} = \kappa_2^{-1}\frac{\|\mathbf{e}\|_2}{\|\mathbf{b}\|_2} \quad \text{when } \mathbf{b} = \beta\mathbf{U}_{*n} \text{ and } \mathbf{e} = \epsilon\mathbf{U}_{*1}.$$

Therefore, if $\mathbf{A}$ is well conditioned, then relatively small uncertainties in $\mathbf{b}$ can't produce relatively large uncertainties in $\mathbf{x}$. But when $\mathbf{A}$ is ill conditioned, it's possible for relatively small uncertainties in $\mathbf{b}$ to have relatively large effects on $\mathbf{x}$, and it's also possible for large uncertainties in $\mathbf{b}$ to have almost no effect on $\mathbf{x}$. Since the direction of $\mathbf{e}$ is almost always unknown, we must guard against the worst case and proceed with caution when dealing with ill-conditioned matrices.

Problem: What if there are uncertainties in both sides of $\mathbf{Ax} = \mathbf{b}$?

Solution: Use calculus to analyze the situation by considering the entries of $\mathbf{A} = \mathbf{A}(t)$ and $\mathbf{b} = \mathbf{b}(t)$ to be differentiable functions of a variable t, and compute the relative size of the derivative of $\mathbf{x} = \mathbf{x}(t)$ by differentiating $\mathbf{b} = \mathbf{Ax}$ to obtain $\mathbf{b}' = (\mathbf{Ax})' = \mathbf{A}'\mathbf{x} + \mathbf{Ax}'$ (with $\star'$ denoting $d\star/dt$), so

$$\begin{aligned}\|\mathbf{x}'\| &= \left\|\mathbf{A}^{-1}\mathbf{b}' - \mathbf{A}^{-1}\mathbf{A}'\mathbf{x}\right\| \le \left\|\mathbf{A}^{-1}\mathbf{b}'\right\| + \left\|\mathbf{A}^{-1}\mathbf{A}'\mathbf{x}\right\| \\ &\le \left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}'\| + \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}'\| \|\mathbf{x}\|.\end{aligned}$$

Consequently,

$$\begin{aligned}\frac{\|\mathbf{x}'\|}{\|\mathbf{x}\|} &\le \frac{\left\|\mathbf{A}^{-1}\right\| \|\mathbf{b}'\|}{\|\mathbf{x}\|} + \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}'\| \\ &\le \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{b}'\|}{\|\mathbf{A}\| \|\mathbf{x}\|} + \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|} \\ &\le \kappa \frac{\|\mathbf{b}'\|}{\|\mathbf{b}\|} + \kappa \frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|} = \kappa \left(\frac{\|\mathbf{b}'\|}{\|\mathbf{b}\|} + \frac{\|\mathbf{A}'\|}{\|\mathbf{A}\|}\right).\end{aligned}$$

In other words, the relative sensitivity of the solution is the sum of the relative sensitivities of $\mathbf{A}$ and $\mathbf{b}$ magnified by $\kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|$. A discrete analog of the above inequality is developed in Exercise 5.12.12.

Conclusion: In all cases, the credibility of the solution to $\mathbf{Ax} = \mathbf{b}$ in the face of uncertainties must be gauged in relation to the condition of $\mathbf{A}$.

As the next example shows, the condition number is pivotal also in determining whether or not the residual $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ is a reliable indicator of the accuracy of an approximate solution $\tilde{\mathbf{x}}$.

Example 5.12.2

Checking an Answer. Suppose that $\tilde{\mathbf{x}}$ is a computed (or otherwise approximate) solution for a nonsingular system $\mathbf{Ax} = \mathbf{b}$, and suppose the accuracy of $\tilde{\mathbf{x}}$ is "checked" by computing the ***residual*** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$. If $\mathbf{r} = \mathbf{0}$, *exactly,* then $\tilde{\mathbf{x}}$ must be the exact solution. But if $\mathbf{r}$ is not exactly zero—say, $\|\mathbf{r}\|_2$ is zero to t significant digits—are we guaranteed that $\tilde{\mathbf{x}}$ is accurate to roughly t significant figures? This question was briefly examined in Example 1.6.3, but it's worth another look.

Problem: To what extent does the size of the residual reflect the accuracy of an approximate solution?

Solution: Without realizing it, we answered this question in Example 5.12.1. To bound the accuracy of $\tilde{\mathbf{x}}$ relative to the exact solution $\mathbf{x}$, write $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ as $\mathbf{A}\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{r}$, and apply (5.12.8) with $\mathbf{e} = \mathbf{r}$ to obtain

$$\kappa^{-1}\frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2} \le \frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \kappa\frac{\|\mathbf{r}\|_2}{\|\mathbf{b}\|_2}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|_2\left\|\mathbf{A}^{-1}\right\|_2. \tag{5.12.9}$$

Therefore, for a well-conditioned $\mathbf{A}$, the residual $\mathbf{r}$ is relatively small if and only if $\tilde{\mathbf{x}}$ is relatively accurate. However, as demonstrated in Example 5.12.1, equality on either side of (5.12.9) is possible, so, when $\mathbf{A}$ is ill conditioned, a very inaccurate approximation $\tilde{\mathbf{x}}$ can produce a small residual $\mathbf{r}$, and a very accurate approximation can produce a large residual.

Conclusion: Residuals are reliable indicators of accuracy only when $\mathbf{A}$ is well conditioned—if $\mathbf{A}$ is ill conditioned, residuals are nearly meaningless.

In addition to measuring the distortion of the unit sphere and gauging the sensitivity of linear systems, singular values provide a measure of how close $\mathbf{A}$ is to a matrix of lower rank.

Distance to Lower-Rank Matrices

If $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ are the nonzero singular values of $\mathbf{A}_{m\times n}$, then for each $k < r$, the distance from $\mathbf{A}$ to the closest matrix of rank k is

$$\sigma_{k+1} = \min_{rank(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_2. \tag{5.12.10}$$

Proof. Suppose $rank\,(\mathbf{B}_{m\times n}) = k$, and let $\mathbf{A} = \mathbf{U}\begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T$ be an SVD for $\mathbf{A}$ with $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$. Define $\mathbf{S} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{k+1})$, and partition $\mathbf{V} = \left(\mathbf{F}_{n\times k+1} \,|\, \mathbf{G}\right)$. Since $rank\,(\mathbf{BF}) \le rank\,(\mathbf{B}) = k$ (by (4.5.2)), $\dim N(\mathbf{BF}) = k+1-rank\,(\mathbf{BF}) \ge 1$, so there is an $\mathbf{x} \in N(\mathbf{BF})$ with $\|\mathbf{x}\|_2 = 1$. Consequently, $\mathbf{BFx} = \mathbf{0}$ and

$$\mathbf{AFx} = \mathbf{U}\begin{pmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T\mathbf{Fx} = \mathbf{U}\begin{pmatrix}\mathbf{S} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \star & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix} = \mathbf{U}\begin{pmatrix}\mathbf{Sx}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}.$$

Since $\|\mathbf{A}-\mathbf{B}\|_2 = \max_{\|\mathbf{y}\|_2=1}\|(\mathbf{A}-\mathbf{B})\mathbf{y}\|_2$, and since $\|\mathbf{Fx}\|_2 = \|\mathbf{x}\|_2 = 1$ (recall (5.2.4), p. 280, and (5.2.13), p. 283),

$$\|\mathbf{A}-\mathbf{B}\|_2^2 \ge \|(\mathbf{A}-\mathbf{B})\mathbf{Fx}\|_2^2 = \|\mathbf{Sx}\|_2^2 = \sum_{i=1}^{k+1}\sigma_i^2 x_i^2 \ge \sigma_{k+1}^2\sum_{i=1}^{k+1}x_i^2 = \sigma_{k+1}^2.$$

Equality holds for $\mathbf{B}_k = \mathbf{U}\begin{pmatrix}\mathbf{D}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T$ with $\mathbf{D}_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, and thus (5.12.10) is proven. ∎

Example 5.12.3

Filtering Noisy Data. The SVD can be a useful tool in applications involving the need to sort through noisy data and lift out relevant information. Suppose that $\mathbf{A}_{m\times n}$ is a matrix containing data that are contaminated with a certain level of noise—e.g., the entries $\mathbf{A}$ might be digital samples of a noisy video or audio signal such as that in Example 5.8.3 (p. 359). The SVD resolves the data in $\mathbf{A}$ into r mutually orthogonal components by writing

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r} \sigma_i \mathbf{Z}_i, \tag{5.12.11}$$

where $\mathbf{Z}_i = \mathbf{u}_i\mathbf{v}_i^T$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. The matrices $\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_r\}$ constitute an orthonormal set because

$$\langle \mathbf{Z}_i | \mathbf{Z}_j \rangle = trace\left(\mathbf{Z}_i^{\mathbf{T}}\mathbf{Z}_j\right) = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

In other words, the SVD (5.12.11) can be regarded as a Fourier expansion as described on p. 299 and, consequently, $\sigma_i = \langle \mathbf{Z}_i | \mathbf{A} \rangle$ can be interpreted as the proportion of $\mathbf{A}$ lying in the "direction" of $\mathbf{Z}_i$. In many applications the noise contamination in $\mathbf{A}$ is random (or nondirectional) in the sense that the noise is distributed more or less uniformly across the $\mathbf{Z}_i$'s. That is, there is about as much noise in the "direction" of one $\mathbf{Z}_i$ as there is in the "direction" of any other. Consequently, we expect each term $\sigma_i\mathbf{Z}_i$ to contain approximately the same level of noise. This means that if $\text{SNR}(\sigma_i\mathbf{Z}_i)$ denotes the *signal-to-noise ratio* in $\sigma_i\mathbf{Z}_i$, then

$$\text{SNR}(\sigma_1\mathbf{Z}_1) \geq \text{SNR}(\sigma_2\mathbf{Z}_2) \geq \cdots \geq \text{SNR}(\sigma_r\mathbf{Z}_r),$$

more or less. If some of the singular values, say, $\sigma_{k+1}, \ldots, \sigma_r$, are small relative to (total noise)$/r$, then the terms $\sigma_{k+1}\mathbf{Z}_{k+1}, \ldots, \sigma_r\mathbf{Z}_r$ have small signal-to-noise ratios. Therefore, if we delete these terms from (5.12.11), then we lose a small part of the total signal, but we remove a disproportionately large component of the total noise in $\mathbf{A}$. This explains why a *truncated* SVD $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i\mathbf{Z}_i$ can, in many instances, filter out some of the noise without losing significant information about the signal in $\mathbf{A}$. Determining the best value of k often requires empirical techniques that vary from application to application, but looking for obvious gaps between large and small singular values is usually a good place to start. The next example presents an interesting application of this idea to building an Internet search engine.

Example 5.12.4

Search Engines. The filtering idea presented in Example 5.12.3 is widely used, but a particularly novel application is the method of *latent semantic indexing* used in the areas of information retrieval and text mining. You can think of this in terms of building an Internet search engine. Start with a dictionary of terms $T_1, T_2, \ldots, T_m$. Terms are usually single words, but sometimes a term may contain more that one word such as "landing gear." It's up to you to decide how extensive your dictionary should be, but even if you use the entire English language, you probably won't be using more than a few hundred-thousand terms, and this is within the capacity of existing computer technology. Each document (or web page) D_j of interest is scanned for key terms (this is called *indexing* the document), and an associated *document vector* $\mathbf{d}_j = (\text{freq}_{1j}, \text{freq}_{2j}, \ldots, \text{freq}_{mj})^T$ is created in which

$$\text{freq}_{ij} = \text{number of times term } T_i \text{ occurs in document } D_j.$$

(More sophisticated search engines use weighted frequency strategies.) After a collection of documents $D_1, D_2, \ldots, D_n$ has been indexed, the associated document vectors $\mathbf{d}_j$ are placed as columns in a *term-by-document matrix*

$$\mathbf{A}_{m\times n} = \big(\mathbf{d}_1 \,|\, \mathbf{d}_2 \,\cdots\, |\, \mathbf{d}_n\big) = \begin{array}{c} \\ T_1 \\ T_2 \\ \vdots \\ T_m \end{array} \begin{array}{c} \begin{array}{cccc} D_1 & D_2 & \cdots & D_n \end{array} \\ \begin{pmatrix} \text{freq}_{11} & \text{freq}_{12} & \cdots & \text{freq}_{1n} \\ \text{freq}_{21} & \text{freq}_{22} & \cdots & \text{freq}_{2n} \\ \vdots & \vdots & & \vdots \\ \text{freq}_{m1} & \text{freq}_{m2} & \cdots & \text{freq}_{mn} \end{pmatrix} \end{array}.$$

Naturally, most entries in each document vector $\mathbf{d}_j$ will be zero, so $\mathbf{A}$ is a sparse matrix—this is good because it means that sparse matrix technology can be applied. When a query composed of a few terms is submitted to the search engine, a *query vector* $\mathbf{q}^T = (q_1, q_2, \ldots, q_n)$ is formed in which

$$q_i = \begin{cases} 1 & \text{if term } T_i \text{ appears in the query,} \\ 0 & \text{otherwise.} \end{cases}$$

(The q_i's might also be weighted.) To measure how well a query $\mathbf{q}$ matches a document D_j, we check how close $\mathbf{q}$ is to $\mathbf{d}_j$ by computing the magnitude of

$$\cos\theta_j = \frac{\mathbf{q}^T\mathbf{d}_j}{\|\mathbf{q}\|_2\,\|\mathbf{d}_j\|_2} = \frac{\mathbf{q}^T\mathbf{A}\mathbf{e}_j}{\|\mathbf{q}\|_2\,\|\mathbf{A}\mathbf{e}_j\|_2}. \tag{5.12.12}$$

If $|\cos\theta_j| \geq \tau$ for some threshold tolerance τ, then document D_j is considered relevant and is returned to the user. Selecting τ is part art and part science that's based on experimentation and desired performance criteria. If the columns of $\mathbf{A}$ along with $\mathbf{q}$ are initially normalized to have unit length, then

$|\mathbf{q}^T\mathbf{A}| = \big(|\cos\theta_1|, |\cos\theta_2|, \ldots, |\cos\theta_n|\big)$ provides the information that allows the search engine to rank the relevance of each document relative to the query. However, due to things like variation and ambiguity in the use of vocabulary, presentation style, and even the indexing process, there is a lot of "noise" in $\mathbf{A}$, so the results in $|\mathbf{q}^T\mathbf{A}|$ are nowhere near being an exact measure of how well query $\mathbf{q}$ matches the various documents. To filter out some of this noise, the techniques of Example 5.12.3 are employed. An SVD $\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is judiciously truncated, and

$$\mathbf{A}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T = \big(\mathbf{u}_1 \mid \cdots \mid \mathbf{u}_k\big) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{pmatrix} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

is used in place of $\mathbf{A}$ in (5.12.12). In other words, instead of using $\cos\theta_j$, query $\mathbf{q}$ is compared with document D_j by using the magnitude of

$$\cos\phi_j = \frac{\mathbf{q}^T\mathbf{A}_k\mathbf{e}_j}{\|\mathbf{q}\|_2\,\|\mathbf{A}_k\mathbf{e}_j\|_2}.$$

To make this more suitable for computation, set $\mathbf{S}_k = \mathbf{D}_k\mathbf{V}_k^T = \big(\mathbf{s}_1 \mid \mathbf{s}_2 \mid \cdots \mid \mathbf{s}_k\big)$, and use

$$\|\mathbf{A}_k\mathbf{e}_j\|_2 = \big\|\mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T\mathbf{e}_j\big\|_2 = \|\mathbf{U}_k\mathbf{s}_j\|_2 = \|\mathbf{s}_j\|_2$$

to write

$$\cos\phi_j = \frac{\mathbf{q}^T\mathbf{U}_k\mathbf{s}_j}{\|\mathbf{q}\|_2\,\|\mathbf{s}_j\|_2}. \tag{5.12.13}$$

The vectors in $\mathbf{U}_k$ and $\mathbf{S}_k$ only need to be computed once (and they can be determined without computing the entire SVD), so (5.12.13) requires very little computation to process each new query. Furthermore, we can be generous in the number of SVD components that are dropped because variation in the use of vocabulary and the ambiguity of many words produces significant noise in $\mathbf{A}$. Coupling this with the fact that numerical accuracy is not an important issue (knowing a cosine to two or three significant digits is sufficient) means that we are more than happy to replace the SVD of $\mathbf{A}$ by a low-rank truncation $\mathbf{A}_k$, where k is *significantly* less than r.

Alternate Query Matching Strategy. An alternate way to measuring how close a given query $\mathbf{q}$ is to a document vector $\mathbf{d}_j$ is to replace the query vector $\mathbf{q}$ in (5.12.12) by the *projected query* $\widetilde{\mathbf{q}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{q}$, where $\mathbf{P}_{R(\mathbf{A})} = \mathbf{U}_r\mathbf{U}_r^T$ is the orthogonal projector onto $R(\mathbf{A})$ along $R(\mathbf{A})^{\perp}$ (Exercise 5.12.15) to produce

$$\cos\widetilde{\theta}_j = \frac{\widetilde{\mathbf{q}}^T\mathbf{A}\mathbf{e}_j}{\|\widetilde{\mathbf{q}}\|_2\,\|\mathbf{A}\mathbf{e}_j\|_2}. \tag{5.12.14}$$

It's proven on p. 435 that $\widetilde{\mathbf{q}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{q}$ is the vector in $R(\mathbf{A})$ (the *document space*) that is closest to $\mathbf{q}$, so using $\widetilde{\mathbf{q}}$ in place of $\mathbf{q}$ has the effect of using the best approximation to $\mathbf{q}$ that is a linear combination of the document vectors $\mathbf{d}_i$. Since $\widetilde{\mathbf{q}}^T\mathbf{A} = \mathbf{q}^T\mathbf{A}$ and $\|\widetilde{\mathbf{q}}\|_2 \le \|\mathbf{q}\|_2$, it follows that $\cos\widetilde{\theta}_j \ge \cos\theta_j$, so more documents are deemed relevant when the projected query is used. Just as in the unprojected query matching strategy, the noise is filtered out by replacing $\mathbf{A}$ in (5.12.14) with a truncated SVD $\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. The result is

$$\cos\widetilde{\phi}_j = \frac{\mathbf{q}^T\mathbf{U}_k\mathbf{s}_j}{\left\|\mathbf{U}_k^T\mathbf{q}\right\|_2 \|\mathbf{s}_j\|_2}$$

and, just as in (5.12.13), $\cos\widetilde{\phi}_j$ is easily and quickly computed for each new query $\mathbf{q}$ because $\mathbf{U}_k$ and $\mathbf{s}_j$ need only be computed once.

The next example shows why singular values are the primary mechanism for numerically determining the rank of a matrix.

Example 5.12.5

Perturbations and Numerical Rank. For $\mathbf{A} \in \Re^{m\times n}$ with $p = \min\{m, n\}$, let $\{\sigma_1, \sigma_2, \ldots, \sigma_p\}$ and $\{\beta_1, \beta_2, \ldots, \beta_p\}$ be all singular values (nonzero as well as any zero ones) for $\mathbf{A}$ and $\mathbf{A} + \mathbf{E}$, respectively.

Problem: Prove that

$$|\sigma_k - \beta_k| \le \|\mathbf{E}\|_2 \quad \text{for each} \quad k = 1, 2, \ldots, p. \tag{5.12.15}$$

Solution: If the SVD for $\mathbf{A}$ given in (5.12.2) is written in the form

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad \text{and if we set} \quad \mathbf{A}_{k-1} = \sum_{i=1}^{k-1} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

then

$$\begin{aligned}
\sigma_k &= \|\mathbf{A} - \mathbf{A}_{k-1}\|_2 = \|\mathbf{A} + \mathbf{E} - \mathbf{A}_{k-1} - \mathbf{E}\|_2 \\
&\ge \|\mathbf{A} + \mathbf{E} - \mathbf{A}_{k-1}\|_2 - \|\mathbf{E}\|_2 \quad \text{(recall (5.1.6) on p. 273)} \\
&\ge \beta_k - \|\mathbf{E}\|_2 \quad \text{by (5.12.10).}
\end{aligned}$$

Couple this with the observation that

$$\begin{aligned}
\sigma_k &= \min_{rank(\mathbf{B})=k-1} \|\mathbf{A} - \mathbf{B}\|_2 = \min_{rank(\mathbf{B})=k-1} \|\mathbf{A} + \mathbf{E} - \mathbf{B} - \mathbf{E}\|_2 \\
&\le \min_{rank(\mathbf{B})=k-1} \|\mathbf{A} + \mathbf{E} - \mathbf{B}\|_2 + \|\mathbf{E}\|_2 = \beta_k + \|\mathbf{E}\|_2
\end{aligned}$$

to conclude that $|\sigma_k - \beta_k| \le \|\mathbf{E}\|_2$.

Problem: Explain why this means that computing the singular values of $\mathbf{A}$ with any stable algorithm (one that returns the exact singular values β_k of a nearby matrix $\mathbf{A}+\mathbf{E}$) is a good way to compute $rank\,(\mathbf{A})$.

Solution: If $rank\,(\mathbf{A}) = r$, then $p-r$ of the σ_k's are exactly zero, so the perturbation result (5.12.15) guarantees that $p-r$ of the computed β_k's cannot be larger than $\|\mathbf{E}\|_2$. So if

$$\beta_1 \geq \cdots \geq \beta_{\tilde{r}} > \|\mathbf{E}\|_2 \geq \beta_{\tilde{r}+1} \geq \cdots \geq \beta_p,$$

then it's reasonable to consider $\tilde{r}$ to be the ***numerical rank*** of $\mathbf{A}$. For most algorithms, $\|\mathbf{E}\|_2$ is not known exactly, but adequate estimates of $\|\mathbf{E}\|_2$ often can be derived. Considerable effort has gone into the development of stable algorithms for computing singular values, but such algorithms are too involved to discuss here—consult an advanced book on matrix computations. Generally speaking, good SVD algorithms have $\|\mathbf{E}\|_2 \approx 5 \times 10^{-t}\|\mathbf{A}\|_2$ when t-digit floating-point arithmetic is used.

Just as the range-nullspace decomposition was used in Example 5.10.5 to define the Drazin inverse of a square matrix, a URV factorization or an SVD can be used to define a generalized inverse for rectangular matrices. For a URV factorization

$$\mathbf{A}_{m\times n} = \mathbf{U}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n}\mathbf{V}^T, \quad \text{we define} \quad \mathbf{A}^{\dagger}_{n\times m} = \mathbf{V}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{n\times m}\mathbf{U}^T$$

to be the ***Moore–Penrose inverse*** (or the ***pseudoinverse***) of $\mathbf{A}$. (Replace $(\star)^T$ by $(\star)^*$ when $\mathbf{A} \in \mathcal{C}^{m\times n}$.) Although the URV factors are not uniquely defined by $\mathbf{A}$, it can be proven that $\mathbf{A}^{\dagger}$ is unique by arguing that $\mathbf{A}^{\dagger}$ is the unique solution to the four Penrose equations

$$\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}, \qquad \mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger},$$

$$\left(\mathbf{A}\mathbf{A}^{\dagger}\right)^T = \mathbf{A}\mathbf{A}^{\dagger}, \qquad \left(\mathbf{A}^{\dagger}\mathbf{A}\right)^T = \mathbf{A}^{\dagger}\mathbf{A},$$

so $\mathbf{A}^{\dagger}$ is the same matrix defined in Exercise 4.5.20. Since it doesn't matter which URV factorization is used, we can use the SVD (5.12.2), in which case $\mathbf{C} = \mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. Some "inverselike" properties that relate $\mathbf{A}^{\dagger}$ to solutions and least squares solutions for linear systems are given in the following summary. Other useful properties appear in the exercises.

Moore–Penrose Pseudoinverse

- In terms of URV factors, the Moore–Penrose pseudoinverse of

$$\mathbf{A}_{m\times n} = \mathbf{U}\begin{pmatrix}\mathbf{C}_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T \quad \text{is} \quad \mathbf{A}^\dagger_{n\times m} = \mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T. \tag{5.12.16}$$

- When $\mathbf{Ax} = \mathbf{b}$ is consistent, $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$ is the solution of minimal euclidean norm. (5.12.17)
- When $\mathbf{Ax} = \mathbf{b}$ is inconsistent, $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$ is the least squares solution of minimal euclidean norm. (5.12.18)
- When an SVD is used, $\mathbf{C} = \mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, so

$$\mathbf{A}^\dagger = \mathbf{V}\begin{pmatrix}\mathbf{D}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T = \sum_{i=1}^{r}\frac{\mathbf{v}_i\mathbf{u}_i^T}{\sigma_i} \quad \text{and} \quad \mathbf{A}^\dagger\mathbf{b} = \sum_{i=1}^{r}\frac{(\mathbf{u}_i^T\mathbf{b})}{\sigma_i}\mathbf{v}_i.$$

Proof. To prove (5.12.17), suppose $\mathbf{Ax}_0 = \mathbf{b}$, and replace $\mathbf{A}$ by $\mathbf{AA}^\dagger\mathbf{A}$ to write $\mathbf{b} = \mathbf{Ax}_0 = \mathbf{AA}^\dagger\mathbf{Ax}_0 = \mathbf{AA}^\dagger\mathbf{b}$. Thus $\mathbf{A}^\dagger\mathbf{b}$ solves $\mathbf{Ax} = \mathbf{b}$ when it is consistent. To see that $\mathbf{A}^\dagger\mathbf{b}$ is the solution of minimal norm, observe that the general solution is $\mathbf{A}^\dagger\mathbf{b} + N(\mathbf{A})$ (a particular solution plus the general solution of the homogeneous equation), so every solution has the form $\mathbf{z} = \mathbf{A}^\dagger\mathbf{b} + \mathbf{n}$, where $\mathbf{n} \in N(\mathbf{A})$. It's not difficult to see that $\mathbf{A}^\dagger\mathbf{b} \in R(\mathbf{A}^\dagger) = R(\mathbf{A}^T)$ (Exercise 5.12.16), so $\mathbf{A}^\dagger\mathbf{b} \perp \mathbf{n}$. Therefore, by the Pythagorean theorem (Exercise 5.4.14),

$$\|\mathbf{z}\|_2^2 = \left\|\mathbf{A}^\dagger\mathbf{b} + \mathbf{n}\right\|_2^2 = \left\|\mathbf{A}^\dagger\mathbf{b}\right\|_2^2 + \|\mathbf{n}\|_2^2 \geq \left\|\mathbf{A}^\dagger\mathbf{b}\right\|_2^2.$$

Equality is possible if and only if $\mathbf{n} = \mathbf{0}$, so $\mathbf{A}^\dagger\mathbf{b}$ is the *unique* minimum norm solution. When $\mathbf{Ax} = \mathbf{b}$ is inconsistent, the least squares solutions are the solutions of the normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$, and it's straightforward to verify that $\mathbf{A}^\dagger\mathbf{b}$ is one such solution (Exercise 5.12.16(c)). To prove that $\mathbf{A}^\dagger\mathbf{b}$ is the least squares solution of minimal norm, apply the same argument used in the consistent case to the normal equations. ∎

Caution! Generalized inverses are useful in formulating theoretical statements such as those above, but, just as in the case of the ordinary inverse, generalized inverses are not practical computational tools. In addition to being computationally inefficient, serious numerical problems result from the fact that $\mathbf{A}^\dagger$ need

not be a continuous function of the entries of $\mathbf{A}$. For example,

$$\mathbf{A}(x) = \begin{pmatrix} 1 & 0 \\ 0 & x \end{pmatrix} \implies \mathbf{A}^\dagger(x) = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1/x \end{pmatrix} & \text{for } x \neq 0, \\ \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} & \text{for } x = 0. \end{cases}$$

Not only is $\mathbf{A}^\dagger(x)$ discontinuous in the sense that $\lim_{x\to 0} \mathbf{A}^\dagger(x) \neq \mathbf{A}^\dagger(0)$, but it is discontinuous in the worst way because as $\mathbf{A}(x)$ comes closer to $\mathbf{A}(0)$ the matrix $\mathbf{A}^\dagger(x)$ moves farther away from $\mathbf{A}^\dagger(0)$. This type of behavior translates into insurmountable computational difficulties because small errors due to round-off (or anything else) can produce enormous errors in the computed $\mathbf{A}^\dagger$, and as errors in $\mathbf{A}$ become smaller the resulting errors in $\mathbf{A}^\dagger$ can become greater. This diabolical fact is also true for the Drazin inverse (p. 399). The inherent numerical problems coupled with the fact that it's extremely rare for an application to require explicit knowledge of the entries of $\mathbf{A}^\dagger$ or $\mathbf{A}^D$ constrains them to being theoretical or notational tools. But don't underestimate this role—go back and read Laplace's statement quoted in the footnote on p. 81.

Example 5.12.6

Another way to view the URV or SVD factorizations in relation to the Moore–Penrose inverse is to consider $\mathbf{A}/_{R(\mathbf{A}^T)}$ and $\mathbf{A}^\dagger/_{R(\mathbf{A})}$, the restrictions of $\mathbf{A}$ and $\mathbf{A}^\dagger$ to $R\left(\mathbf{A}^T\right)$ and $R\left(\mathbf{A}\right)$, respectively. Begin by making the straightforward observations that $R\left(\mathbf{A}^\dagger\right) = R\left(\mathbf{A}^T\right)$ and $N\left(\mathbf{A}^\dagger\right) = N\left(\mathbf{A}^T\right)$ (Exercise 5.12.16). Since $\Re^n = R\left(\mathbf{A}^T\right) \oplus N\left(\mathbf{A}\right)$ and $\Re^m = R\left(\mathbf{A}\right) \oplus N\left(\mathbf{A}^T\right)$, it follows that $R\left(\mathbf{A}\right) = \mathbf{A}(\Re^n) = \mathbf{A}(R\left(\mathbf{A}^T\right))$ and $R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^\dagger\right) = \mathbf{A}^\dagger(\Re^m) = \mathbf{A}^\dagger(R\left(\mathbf{A}\right))$. In other words, $\mathbf{A}/_{R(\mathbf{A}^T)}$ and $\mathbf{A}^\dagger/_{R(\mathbf{A})}$ are linear transformations such that

$$\mathbf{A}/_{R(\mathbf{A}^T)} : R\left(\mathbf{A}^T\right) \to R\left(\mathbf{A}\right) \quad \text{and} \quad \mathbf{A}^\dagger/_{R(\mathbf{A})} : R\left(\mathbf{A}\right) \to R\left(\mathbf{A}^T\right).$$

If $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ and $\mathcal{B}' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ are the first r columns from $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$ and $\mathbf{V} = \left(\mathbf{V}_1 \,|\, \mathbf{V}_2\right)$ in (5.11.11), then $\mathbf{A}\mathbf{V}_1 = \mathbf{U}_1\mathbf{C}$ and $\mathbf{A}^\dagger\mathbf{U}_1 = \mathbf{V}_1\mathbf{C}^{-1}$ implies (recall (4.7.4)) that

$$\left[\mathbf{A}/_{R(\mathbf{A}^T)}\right]_{\mathcal{B}'\mathcal{B}} = \mathbf{C} \quad \text{and} \quad \left[\mathbf{A}^\dagger/_{R(\mathbf{A})}\right]_{\mathcal{B}\mathcal{B}'} = \mathbf{C}^{-1}. \tag{5.12.19}$$

If left-hand and right-hand singular vectors from the SVD (5.12.2) are used in $\mathcal{B}$ and $\mathcal{B}'$, respectively, then $\mathbf{C} = \mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. Thus (5.12.19) reveals the exact sense in which $\mathbf{A}$ and $\mathbf{A}^\dagger$ are "inverses." Compare these results with the analogous statements for the Drazin inverse in Example 5.10.5 on p. 399.

Exercises for section 5.12

5.12.1. Following the derivation in the text, find an SVD for

$$\mathbf{C} = \begin{pmatrix} -4 & -6 \\ 3 & -8 \end{pmatrix}.$$

5.12.2. If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of $\mathbf{A}$, then it can be shown that the function $\nu_k(\mathbf{A}) = \left(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2\right)^{1/2}$ defines a unitarily invariant norm (recall Exercise 5.6.9) for $\Re^{m\times n}$ (or $\mathcal{C}^{m\times n}$) for each $k = 1, 2, \ldots, r$. Explain why the 2-norm and the Frobenius norm (p. 279) are the extreme cases in the sense that $\|\mathbf{A}\|_2^2 = \sigma_1^2$ and $\|\mathbf{A}\|_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$.

5.12.3. Each of the four common matrix norms can be bounded above and below by a constant multiple of each of the other matrix norms. To be precise, $\|\mathbf{A}\|_i \leq \alpha \|\mathbf{A}\|_j$, where α is the (i,j)-entry in the following matrix.

$$\begin{array}{c} \\ 1 \\ 2 \\ \infty \\ F \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \infty & F \end{array} \\ \begin{pmatrix} * & \sqrt{n} & n & \sqrt{n} \\ \sqrt{n} & * & \sqrt{n} & 1 \\ n & \sqrt{n} & * & \sqrt{n} \\ \sqrt{n} & \sqrt{n} & \sqrt{n} & * \end{pmatrix} \end{array}.$$

For analyzing limiting behavior, it therefore makes no difference which of these norms is used, so they are said to be *equivalent matrix norms.* (A similar statement for vector norms was given in Exercise 5.1.8.) Explain why the $(2, F)$ and the $(F, 2)$ entries are correct.

5.12.4. Prove that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of a rank r matrix $\mathbf{A}$, and if $\|\mathbf{E}\|_2 < \sigma_r$, then $rank\,(\mathbf{A} + \mathbf{E}) \geq rank\,(\mathbf{A})$. **Note:** This clarifies the meaning of the term "sufficiently small" in the assertion on p. 216 that small perturbations can't reduce rank.

5.12.5. **Image of the Unit Sphere.** Extend the result on p. 414 concerning the image of the unit sphere to include singular and rectangular matrices by showing that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the nonzero singular values of $\mathbf{A}_{m\times n}$, then the image $\mathbf{A}(\mathcal{S}_2) \subset \Re^m$ of the unit 2-sphere $\mathcal{S}_2 \subset \Re^n$ is an ellipsoid (possibly degenerate) in which the kth semiaxis is $\sigma_k \mathbf{U}_{*k} = \mathbf{A}\mathbf{V}_{*k}$, where $\mathbf{U}_{*k}$ and $\mathbf{V}_{*k}$ are respective left-hand and right-hand singular vectors for $\mathbf{A}$.

5.12.6. Prove that if σ_r is the smallest nonzero singular value of $\mathbf{A}_{m\times n}$, then

$$\sigma_r = \min_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in R(\mathbf{A}^T)}} \|\mathbf{A}\mathbf{x}\|_2 = 1/\left\|\mathbf{A}^\dagger\right\|_2,$$

which is the generalization of (5.12.5).

5.12.7. **Generalized Condition Number.** Extend the bound in (5.12.8) to include singular and rectangular matrices by showing that if $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the respective minimum 2-norm solutions of consistent systems $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} = \mathbf{b} - \mathbf{e}$, then

$$\kappa^{-1}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} \le \frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \kappa\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\|\left\|\mathbf{A}^\dagger\right\|.$$

Can the same reasoning given in Example 5.12.1 be used to argue that for $\|\star\|_2$, the upper and lower bounds are attainable for every $\mathbf{A}$?

5.12.8. Prove that if $|\epsilon| < \sigma_r^2$ for the smallest nonzero singular value of $\mathbf{A}_{m\times n}$, then $(\mathbf{A}^T\mathbf{A} + \epsilon\mathbf{I})^{-1}$ exists, and $\lim_{\epsilon\to 0}(\mathbf{A}^T\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{A}^T = \mathbf{A}^\dagger$.

5.12.9. Consider a system $\mathbf{A}\mathbf{x} = \mathbf{b}$ in which

$$\mathbf{A} = \begin{pmatrix} .835 & .667 \\ .333 & .266 \end{pmatrix},$$

and suppose $\mathbf{b}$ is subject to an uncertainty $\mathbf{e}$. Using ∞-norms, determine the directions of $\mathbf{b}$ and $\mathbf{e}$ that give rise to the worst-case scenario in (5.12.8) in the sense that $\|\mathbf{x}-\tilde{\mathbf{x}}\|_\infty / \|\mathbf{x}\|_\infty = \kappa_\infty\|\mathbf{e}\|_\infty / \|\mathbf{b}\|_\infty$.

5.12.10. An ill-conditioned matrix is suspected when a small pivot u_{ii} emerges during the LU factorization of $\mathbf{A}$ because $\left[\mathbf{U}^{-1}\right]_{ii} = 1/u_{ii}$ is then large, and this opens the possibility of $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ having large entries. Unfortunately, this is not an absolute test, and no guarantees about conditioning can be made from the pivots alone.

(a) Construct an example of a matrix that is well conditioned but has a small pivot.

(b) Construct an example of a matrix that is ill conditioned but has no small pivots.

5.12.11. Bound the relative uncertainty in the solution of a nonsingular system $\mathbf{Ax} = \mathbf{b}$ for which there is some uncertainty in $\mathbf{A}$ but not in $\mathbf{b}$ by showing that if $(\mathbf{A} - \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}$, where $\alpha = \left\|\mathbf{A}^{-1}\mathbf{E}\right\| < 1$ for any matrix norm such that $\|\mathbf{I}\| = 1$, then

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa}{1 - \alpha} \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|}, \quad \text{where} \quad \kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|.$$

Note: If the 2-norm is used, then $\|\mathbf{E}\|_2 < \sigma_n$ insures $\alpha < 1$.
Hint: If $\mathbf{B} = \mathbf{A}^{-1}\mathbf{E}$, then $\mathbf{A} - \mathbf{E} = \mathbf{A}(\mathbf{I} - \mathbf{B})$, and $\alpha = \|\mathbf{B}\| < 1$ $\Longrightarrow$ $\left\|\mathbf{B}^k\right\| \leq \|\mathbf{B}\|^k \to 0$ $\Longrightarrow$ $\mathbf{B}^k \to \mathbf{0}$, so the Neumann series expansion (p. 126) yields $(\mathbf{I} - \mathbf{B})^{-1} = \sum_{i=0}^{\infty} \mathbf{B}^i$.

5.12.12. Now bound the relative uncertainty in the solution of a nonsingular system $\mathbf{Ax} = \mathbf{b}$ for which there is some uncertainty in both $\mathbf{A}$ and $\mathbf{b}$ by showing that if $(\mathbf{A} - \mathbf{E})\tilde{\mathbf{x}} = \mathbf{b} - \mathbf{e}$, where $\alpha = \left\|\mathbf{A}^{-1}\mathbf{E}\right\| < 1$ for any matrix norm such that $\|\mathbf{I}\| = 1$, then

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa}{1 - \kappa \|\mathbf{E}\| / \|\mathbf{A}\|} \left(\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|} + \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} \right), \quad \text{where} \quad \kappa = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|.$$

Note: If the 2-norm is used, then $\|\mathbf{E}\|_2 < \sigma_n$ insures $\alpha < 1$. This exercise underscores the conclusion of Example 5.12.1 stating that if $\mathbf{A}$ is well conditioned, and if the relative uncertainties in $\mathbf{A}$ and $\mathbf{b}$ are small, then the relative uncertainty in $\mathbf{x}$ must be small.

5.12.13. Consider the matrix $\mathbf{A} = \begin{pmatrix} -4 & -2 & -4 & -2 \\ 2 & -2 & 2 & 1 \\ -4 & 1 & -4 & -2 \end{pmatrix}$.

(a) Use the URV factorization you computed in Exercise 5.11.8 to determine $\mathbf{A}^\dagger$.

(b) Now use the URV factorization you obtained in Exercise 5.11.9 to determine $\mathbf{A}^\dagger$. Do your results agree with those of part (a)?

5.12.14. For matrix $\mathbf{A}$ in Exercise 5.11.8, and for $\mathbf{b} = \begin{pmatrix} -12 & 3 & -9 \end{pmatrix}^T$, find the solution of $\mathbf{Ax} = \mathbf{b}$ that has minimum euclidean norm.

5.12.15. Suppose $\mathbf{A} = \mathbf{URV}^T$ is a URV factorization (so it could be an SVD) of an $m \times n$ matrix of rank r, and suppose $\mathbf{U}$ is partitioned as $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$, where $\mathbf{U}_1$ is $m \times r$. Prove that $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T = \mathbf{AA}^\dagger$ is the projector onto $R(\mathbf{A})$ along $N\left(\mathbf{A}^T\right)$. In this case, $\mathbf{P}$ is said to be an *orthogonal projector* because its range is orthogonal to its nullspace. What is the orthogonal projector onto $N\left(\mathbf{A}^T\right)$ along $R(\mathbf{A})$? (Orthogonal projectors are discussed in more detail on p. 429.)

5.12.16. Establish the following properties of $\mathbf{A}^\dagger$.

(a) $\mathbf{A}^\dagger = \mathbf{A}^{-1}$ when $\mathbf{A}$ is nonsingular.

(b) $(\mathbf{A}^\dagger)^\dagger = \mathbf{A}$.

(c) $(\mathbf{A}^\dagger)^T = (\mathbf{A}^T)^\dagger$.

(d) $\mathbf{A}^\dagger = \begin{cases} (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T & \text{when } rank\,(\mathbf{A}_{m\times n}) = n, \\ \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} & \text{when } rank\,(\mathbf{A}_{m\times n}) = m. \end{cases}$

(e) $\mathbf{A}^T = \mathbf{A}^T\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

(f) $\mathbf{A}^\dagger = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^\dagger = (\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

(g) $R\left(\mathbf{A}^\dagger\right) = R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^\dagger\mathbf{A}\right)$, and
$N\left(\mathbf{A}^\dagger\right) = N\left(\mathbf{A}^T\right) = N\left(\mathbf{A}\mathbf{A}^\dagger\right)$.

(h) $(\mathbf{PAQ})^\dagger = \mathbf{Q}^T\mathbf{A}^\dagger\mathbf{P}^T$ when $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices, but in general $(\mathbf{AB})^\dagger \neq \mathbf{B}^\dagger\mathbf{A}^\dagger$ (the reverse-order law fails).

(i) $(\mathbf{A}^T\mathbf{A})^\dagger = \mathbf{A}^\dagger(\mathbf{A}^T)^\dagger$ and $(\mathbf{A}\mathbf{A}^T)^\dagger = (\mathbf{A}^T)^\dagger\mathbf{A}^\dagger$.

5.12.17. Explain why $\mathbf{A}^\dagger = \mathbf{A}^D$ if and only if $\mathbf{A}$ is an RPN matrix.

5.12.18. Let $\mathbf{X}, \mathbf{Y} \in \Re^{m\times n}$ be such that $R(\mathbf{X}) \perp R(\mathbf{Y})$.

(a) Establish the Pythagorean theorem for matrices by proving

$$\|\mathbf{X}+\mathbf{Y}\|_F^2 = \|\mathbf{X}\|_F^2 + \|\mathbf{Y}\|_F^2 .$$

(b) Give an example to show that the result of part (a) does not hold for the matrix 2-norm.

(c) Demonstrate that $\mathbf{A}^\dagger$ is the ***best approximate inverse*** for $\mathbf{A}$ in the sense that $\mathbf{A}^\dagger$ is the matrix of smallest Frobenius norm that minimizes $\|\mathbf{I} - \mathbf{AX}\|_F$.

5.13 ORTHOGONAL PROJECTION

As discussed in §5.9, every pair of complementary subspaces defines a projector. But when the complementary subspaces happen to be *orthogonal* complements, the resulting projector has some particularly nice properties, and the purpose of this section is to develop this special case in more detail. Discussions are in the context of real spaces, but generalizations to complex spaces are straightforward by replacing $(\star)^T$ by $(\star)^*$ and "orthogonal matrix" by "unitary matrix."

If $\mathcal{M}$ is a subspace of an inner-product space $\mathcal{V}$, then $\mathcal{V} = \mathcal{M} \oplus \mathcal{M}^\perp$ by (5.11.1), and each $\mathbf{v} \in \mathcal{V}$ can be written uniquely as $\mathbf{v} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^\perp$ by (5.9.3). The vector $\mathbf{m}$ was defined on p. 385 to be the projection of $\mathbf{v}$ onto $\mathcal{M}$ along $\mathcal{M}^\perp$, so the following definitions are natural.

Orthogonal Projection

For $\mathbf{v} \in \mathcal{V}$, let $\mathbf{v} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in \mathcal{M}$ and $\mathbf{n} \in \mathcal{M}^\perp$.

- $\mathbf{m}$ is called the ***orthogonal projection*** of $\mathbf{v}$ onto $\mathcal{M}$.
- The projector $\mathbf{P}_{\mathcal{M}}$ onto $\mathcal{M}$ along $\mathcal{M}^\perp$ is called the ***orthogonal projector*** onto $\mathcal{M}$.
- $\mathbf{P}_{\mathcal{M}}$ is the unique linear operator such that $\mathbf{P}_{\mathcal{M}}\mathbf{v} = \mathbf{m}$ (see p. 386).

These ideas are illustrated illustrated in Figure 5.13.1 for $\mathcal{V} = \Re^3$.

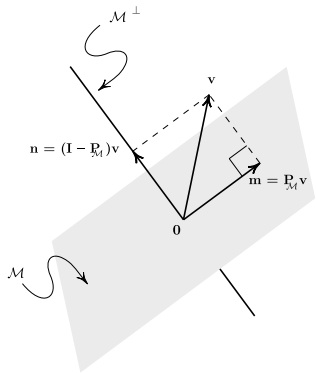


FIGURE 5.13.1

Given an arbitrary pair of complementary subspaces $\mathcal{M}$, $\mathcal{N}$ of $\Re^n$, formula (5.9.12) on p. 386 says that the projector $\mathbf{P}$ onto $\mathcal{M}$ along $\mathcal{N}$ is given by

$$\mathbf{P} = \big(\mathbf{M}\,|\,\mathbf{N}\big)\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\big(\mathbf{M}\,|\,\mathbf{N}\big)^{-1} = \big(\mathbf{M}\,|\,\mathbf{0}\big)\big(\mathbf{M}\,|\,\mathbf{N}\big)^{-1}, \tag{5.13.1}$$

where the columns of $\mathbf{M}$ and $\mathbf{N}$ constitute bases for $\mathcal{M}$ and $\mathcal{N}$, respectively. So, how does this expression simplify when $\mathcal{N} = \mathcal{M}^\perp$? To answer the question,

observe that if $\mathcal{N} = \mathcal{M}^{\perp}$, then $\mathbf{N}^T\mathbf{M} = \mathbf{0}$ and $\mathbf{M}^T\mathbf{N} = \mathbf{0}$. Furthermore, if $\dim \mathcal{M} = r$, then $\mathbf{M}^T\mathbf{M}$ is $r \times r$, and $rank\left(\mathbf{M}^T\mathbf{M}\right) = rank\left(\mathbf{M}\right) = r$ by (4.5.4), so $\mathbf{M}^T\mathbf{M}$ is nonsingular. Therefore, if the columns of $\mathbf{N}$ are chosen to be an orthonormal basis for $\mathcal{M}^{\perp}$, then

$$\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right)\left(\mathbf{M}\,|\,\mathbf{N}\right)=\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} \implies \left(\mathbf{M}\,|\,\mathbf{N}\right)^{-1}=\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right).$$

This together with (5.13.1) says the orthogonal projector onto $\mathcal{M}$ is given by

$$\mathbf{P}_{\mathcal{M}} = \left(\mathbf{M}\,|\,\mathbf{0}\right)\left(\begin{array}{c}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T\\ \hline \mathbf{N}^T\end{array}\right) = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T. \tag{5.13.2}$$

As discussed in §5.9, the projector associated with any given pair of complementary subspaces is unique, and it doesn't matter which bases are used to form $\mathbf{M}$ and $\mathbf{N}$ in (5.13.1). Consequently, formula $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T$ is independent of the choice of $\mathbf{M}$—just as long as its columns constitute some basis for $\mathcal{M}$. In particular, the columns of $\mathbf{M}$ need not be an *orthonormal* basis for $\mathcal{M}$. But if they are, then $\mathbf{M}^T\mathbf{M} = \mathbf{I}$, and (5.13.2) becomes $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\mathbf{M}^T$. Moreover, if the columns of $\mathbf{M}$ and $\mathbf{N}$ constitute orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively, then $\mathbf{U} = \left(\mathbf{M}\,|\,\mathbf{N}\right)$ is an orthogonal matrix, and (5.13.1) becomes

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T.$$

In other words, every orthogonal projector is orthogonally similar to a diagonal matrix in which the diagonal entries are 1's and 0's.

Below is a summary of the formulas used to build orthogonal projectors.

Constructing Orthogonal Projectors

Let $\mathcal{M}$ be an r-dimensional subspace of $\Re^n$, and let the columns of $\mathbf{M}_{n\times r}$ and $\mathbf{N}_{n\times n-r}$ be bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively. The orthogonal projectors onto $\mathcal{M}$ and $\mathcal{M}^{\perp}$ are

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T$. (5.13.3)

If $\mathbf{M}$ and $\mathbf{N}$ contain *orthonormal* bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, then

- $\mathbf{P}_{\mathcal{M}} = \mathbf{M}\mathbf{M}^T$ and $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{N}\mathbf{N}^T$. (5.13.4)
- $\mathbf{P}_{\mathcal{M}} = \mathbf{U}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T$, where $\mathbf{U} = \left(\mathbf{M}\,|\,\mathbf{N}\right)$. (5.13.5)
- $\mathbf{P}_{\mathcal{M}^{\perp}} = \mathbf{I} - \mathbf{P}_{\mathcal{M}}$ in all cases. (5.13.6)

Note: Extensions of (5.13.3) appear on p. 634.

Example 5.13.1

Problem: Let $\mathbf{u}_{n\times 1} \neq \mathbf{0}$, and consider the line $\mathcal{L} = span\,\{\mathbf{u}\}$. Construct the orthogonal projector onto $\mathcal{L}$, and then determine the orthogonal projection of a vector $\mathbf{x}_{n\times 1}$ onto $\mathcal{L}$.

Solution: The vector $\mathbf{u}$ by itself is a basis for $\mathcal{L}$, so, according to (5.13.3),

$$\mathbf{P}_{\mathcal{L}} = \mathbf{u}\left(\mathbf{u}^T\mathbf{u}\right)^{-1}\mathbf{u}^T = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}$$

is the orthogonal projector onto $\mathcal{L}$. The orthogonal projection of a vector $\mathbf{x}$ onto $\mathcal{L}$ is therefore given by

$$\mathbf{P}_{\mathcal{L}}\mathbf{x} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{x} = \left(\frac{\mathbf{u}^T\mathbf{x}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u}.$$

Note: If $\|\mathbf{u}\|_2 = 1$, then $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$, so $\mathbf{P}_{\mathcal{L}}\mathbf{x} = \mathbf{u}\mathbf{u}^T\mathbf{x} = (\mathbf{u}^T\mathbf{x})\mathbf{u}$, and

$$\|\mathbf{P}_{\mathcal{L}}\mathbf{x}\|_2 = |\mathbf{u}^T\mathbf{x}|\ \|\mathbf{u}\|_2 = |\mathbf{u}^T\mathbf{x}|.$$

This yields a geometrical interpretation for the magnitude of the standard inner product. It says that if $\mathbf{u}$ is a vector of unit length in $\mathcal{L}$, then, as illustrated in Figure 5.13.2, $|\mathbf{u}^T\mathbf{x}|$ is the length of the orthogonal projection of $\mathbf{x}$ onto the line spanned by $\mathbf{u}$.

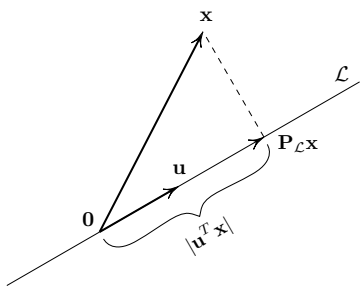


FIGURE 5.13.2

Finally, notice that since $\mathbf{P}_{\mathcal{L}} = \mathbf{u}\mathbf{u}^T$ is the orthogonal projector onto $\mathcal{L}$, it must be the case that $\mathbf{P}_{\mathcal{L}^\perp} = \mathbf{I} - \mathbf{P}_{\mathcal{L}} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$ is the orthogonal projection onto $\mathcal{L}^\perp$. This was called an ***elementary orthogonal projector*** on p. 322—go back and reexamine Figure 5.6.1.

Example 5.13.2

Volume, Gram–Schmidt, and QR. A solid in $\Re^m$ with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is called an n-dimensional ***parallelepiped***. As shown in the shaded portions of Figure 5.13.3, a two-dimensional parallelepiped is a parallelogram, and a three-dimensional parallelepiped is a skewed rectangular box.

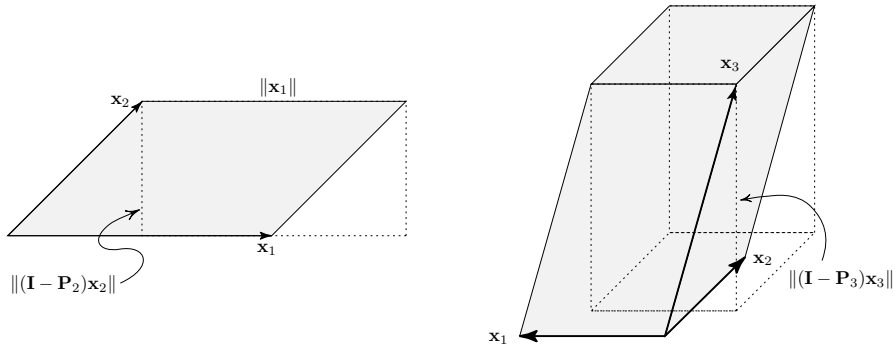


FIGURE 5.13.3

Problem: Determine the volumes of a two-dimensional and a three-dimensional parallelepiped, and then make the natural extension to define the volume of an n-dimensional parallelepiped.

Solution: In the two-dimensional case, volume is area, and it's evident from Figure 5.13.3 that the area of the shaded parallelogram is the same as the area of the dotted rectangle. The width of the dotted rectangle is $\nu_1 = \|\mathbf{x}_1\|_2$, and the height is $\nu_2 = \|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2$, where $\mathbf{P}_2$ is the orthogonal projector onto the space (line) spanned by $\mathbf{x}_1$, and $\mathbf{I}-\mathbf{P}_2$ is the orthogonal projector onto $span\,\{\mathbf{x}_1\}^{\perp}$. In other words, the area, V_2, of the parallelogram is the length of its base times its *projected height*, ν_2, so

$$V_2 = \|\mathbf{x}_1\|_2\,\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2.$$

Similarly, the volume of a three-dimensional parallelepiped is the area of its base times its projected height. The area of the base was just determined to be $V_2 = \|\mathbf{x}_1\|_2\,\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2 = \nu_1\nu_2$, and it's evident from Figure 5.13.3 that the projected height is $\nu_3 = \|(\mathbf{I}-\mathbf{P}_3)\mathbf{x}_3\|_2$, where $\mathbf{P}_3$ is the orthogonal projector onto $span\,\{\mathbf{x}_1,\mathbf{x}_2\}$. Therefore, the volume of the parallelepiped generated by $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$ is

$$V_3 = \|\mathbf{x}_1\|_2\,\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2\,\|(\mathbf{I}-\mathbf{P}_3)\mathbf{x}_3\|_2 = \nu_1\nu_2\nu_3.$$

It's now clear how to inductively define V_4, V_5, etc. In general, the volume of the parallelepiped generated by a linearly independent set $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ is

$$V_n = \|\mathbf{x}_1\|_2\,\|(\mathbf{I}-\mathbf{P}_2)\mathbf{x}_2\|_2\,\|(\mathbf{I}-\mathbf{P}_3)\mathbf{x}_3\|_2\cdots\|(\mathbf{I}-\mathbf{P}_n)\mathbf{x}_n\|_2 = \nu_1\nu_2\cdots\nu_n,$$

where $\mathbf{P}_k$ is the orthogonal projector onto $span\,\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{k-1}\}$, and where

$$\nu_1 = \|\mathbf{x}_1\|_2 \quad \text{and} \quad \nu_k = \|(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_k\|_2 \quad \text{for} \quad k > 1. \tag{5.13.7}$$

Note that if $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$ is an *orthogonal* set, $V_n = \|\mathbf{x}_1\|_2\,\|\mathbf{x}_2\|_2\cdots\|\mathbf{x}_n\|_2$, which is what we would expect.

Connections with Gram–Schmidt and QR. Recall from (5.5.4) on p. 309 that the vectors in the Gram–Schmidt sequence generated from a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \Re^m$ are $\mathbf{u}_1 = \mathbf{x}_1 / \|\mathbf{x}_1\|_2$ and

$$\mathbf{u}_k = \frac{\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k}{\left\|\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k\right\|_2}, \quad \text{where} \quad \mathbf{U}_k = \left[\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_{k-1}\right] \quad \text{for } k > 1.$$

Since $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}\}$ is an *orthonormal* basis for $span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$, it follows from (5.13.4) that $\mathbf{U}_k\mathbf{U}_k^T$ must be the orthogonal projector onto $span\,\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$. Hence $\mathbf{U}_k\mathbf{U}_k^T = \mathbf{P}_k$ and $(\mathbf{I} - \mathbf{P}_k)\mathbf{x}_k = (\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T)\mathbf{x}_k$, so $\left\|\left(\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^T\right)\mathbf{x}_k\right\|_2 = \nu_k$ is the kth projected height in (5.13.7). This means that when the Gram–Schmidt equations are written in the form of a QR factorization as explained on p. 311, the diagonal elements of the upper-triangular matrix $\mathbf{R}$ are the ν_k's. Consequently, the product of the diagonal entries in $\mathbf{R}$ is the volume of the parallelepiped generated by the $\mathbf{x}_k$'s. But the QR factorization of $\mathbf{A} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\right]$ is unique (Exercise 5.5.8), so it doesn't matter whether Gram–Schmidt or another method is used to determine the QR factors. Therefore, we arrive at the following conclusion.

- If $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$ is the (rectangular) QR factorization of a matrix with linearly independent columns, then the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \nu_1\nu_2\cdots\nu_n$, where the ν_k's are the diagonal elements of $\mathbf{R}$. We will see on p. 468 what this means in terms of determinants.

Of course, not all projectors are *orthogonal* projectors, so a natural question to ask is, "What characteristic features distinguish orthogonal projectors from more general oblique projectors?" Some answers are given below.

Orthogonal Projectors

Suppose that $\mathbf{P} \in \Re^{n\times n}$ is a projector—i.e., $\mathbf{P}^2 = \mathbf{P}$. The following statements are equivalent to saying that $\mathbf{P}$ is an *orthogonal* projector.

- $R(\mathbf{P}) \perp N(\mathbf{P})$. (5.13.8)
- $\mathbf{P}^T = \mathbf{P}$ (i.e., orthogonal projector $\Longleftrightarrow \mathbf{P}^2 = \mathbf{P} = \mathbf{P}^T$). (5.13.9)
- $\|\mathbf{P}\|_2 = 1$ for the matrix 2-norm (p. 281). (5.13.10)

Proof. Every projector projects vectors onto its range along (parallel to) its nullspace, so statement (5.13.8) is essentially a restatement of the definition of an orthogonal projector. To prove (5.13.9), note that if $\mathbf{P}$ is an orthogonal projector, then (5.13.3) insures that $\mathbf{P}$ is symmetric. Conversely, if a projector

$\mathbf{P}$ is symmetric, then it must be an orthogonal projector because (5.11.5) on p. 405 allows us to write

$$\mathbf{P} = \mathbf{P}^T \implies R(\mathbf{P}) = R\left(\mathbf{P}^T\right) \implies R(\mathbf{P}) \perp N(\mathbf{P}).$$

To see why (5.13.10) characterizes projectors that are orthogonal, refer back to Example 5.9.2 on p. 389 (or look ahead to (5.15.3)) and note that $\|\mathbf{P}\|_2 = 1/\sin\theta$, where θ is the angle between $R(\mathbf{P})$ and $N(\mathbf{P})$. This makes it clear that $\|\mathbf{P}\|_2 \geq 1$ for all projectors, and $\|\mathbf{P}\|_2 = 1$ if and only if $\theta = \pi/2$, (i.e., if and only if $R(\mathbf{P}) \perp N(\mathbf{P})$). ∎

Example 5.13.3

Problem: For $\mathbf{A} \in \Re^{m\times n}$ such that $rank(\mathbf{A}) = r$, describe the orthogonal projectors onto each of the four fundamental subspaces of $\mathbf{A}$.

Solution 1: Let $\mathbf{B}_{m\times r}$ and $\mathbf{N}_{n\times n-r}$ be matrices whose columns are bases for $R(\mathbf{A})$ and $N(\mathbf{A})$, respectively—e.g., $\mathbf{B}$ might contain the basic columns of $\mathbf{A}$. The orthogonal decomposition theorem on p. 405 says $R(\mathbf{A})^\perp = N\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})^\perp = R\left(\mathbf{A}^T\right)$, so, by making use of (5.13.3) and (5.13.6), we can write

$$\mathbf{P}_{R(\mathbf{A})} = \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T,$$

$$\mathbf{P}_{N(\mathbf{A}^T)} = \mathbf{P}_{R(\mathbf{A})^\perp} = \mathbf{I} - \mathbf{P}_{R(\mathbf{A})} = \mathbf{I} - \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T,$$

$$\mathbf{P}_{N(\mathbf{A})} = \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T,$$

$$\mathbf{P}_{R(\mathbf{A}^T)} = \mathbf{P}_{N(\mathbf{A})^\perp} = \mathbf{I} - \mathbf{P}_{N(\mathbf{A})} = \mathbf{I} - \mathbf{N}\left(\mathbf{N}^T\mathbf{N}\right)^{-1}\mathbf{N}^T.$$

Note: If $rank(\mathbf{A}) = n$, then all columns of $\mathbf{A}$ are basic and

$$\mathbf{P}_{R(\mathbf{A})} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T. \tag{5.13.11}$$

Solution 2: Another way to describe these projectors is to make use of the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ (p. 423). Recall that if $\mathbf{A}$ has a URV factorization

$$\mathbf{A} = \mathbf{U}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T, \quad \text{then} \quad \mathbf{A}^\dagger = \mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T,$$

where $\mathbf{U} = \left(\mathbf{U}_1 \,|\, \mathbf{U}_2\right)$ and $\mathbf{V} = \left(\mathbf{V}_1 \,|\, \mathbf{V}_2\right)$ are orthogonal matrices in which the columns of $\mathbf{U}_1$ and $\mathbf{V}_1$ constitute orthonormal bases for $R(\mathbf{A})$ and $R\left(\mathbf{A}^T\right)$, respectively, and the columns of $\mathbf{U}_2$ and $\mathbf{V}_2$ are orthonormal bases for $N\left(\mathbf{A}^T\right)$ and $N(\mathbf{A})$, respectively. Computing the products $\mathbf{A}\mathbf{A}^\dagger$ and $\mathbf{A}^\dagger\mathbf{A}$ reveals

$$\mathbf{A}\mathbf{A}^\dagger = \mathbf{U}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T = \mathbf{U}_1\mathbf{U}_1^T \quad \text{and} \quad \mathbf{A}^\dagger\mathbf{A} = \mathbf{V}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T = \mathbf{V}_1\mathbf{V}_1^T,$$

so, according to (5.13.4),

$$\begin{aligned} \mathbf{P}_{R(\mathbf{A})} &= \mathbf{U}_1\mathbf{U}_1^T = \mathbf{A}\mathbf{A}^\dagger, & \mathbf{P}_{N(\mathbf{A}^T)} &= \mathbf{I} - \mathbf{P}_{R(\mathbf{A})} = \mathbf{I} - \mathbf{A}\mathbf{A}^\dagger, \\ \mathbf{P}_{R(\mathbf{A}^T)} &= \mathbf{V}_1\mathbf{V}_1^T = \mathbf{A}^\dagger\mathbf{A}, & \mathbf{P}_{N(\mathbf{A})} &= \mathbf{I} - \mathbf{P}_{R(\mathbf{A}^T)} = \mathbf{I} - \mathbf{A}^\dagger\mathbf{A}. \end{aligned} \tag{5.13.12}$$

The notion of orthogonal projection in higher-dimensional spaces is consistent with the visual geometry in $\Re^2$ and $\Re^3$. In particular, it is visually evident from Figure 5.13.4 that if $\mathcal{M}$ is a subspace of $\Re^3$, and if $\mathbf{b}$ is a vector outside of $\mathcal{M}$, then the point in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$.

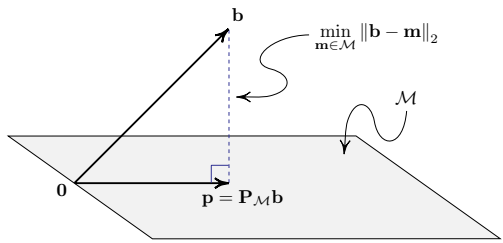


FIGURE 5.13.4

The situation is exactly the same in higher dimensions. But rather than using our eyes to understand why, we use mathematics—it's surprising just how easy it is to "see" such things in abstract spaces.

Closest Point Theorem

Let $\mathcal{M}$ be a subspace of an inner-product space $\mathcal{V}$, and let $\mathbf{b}$ be a vector in $\mathcal{V}$. The unique vector in $\mathcal{M}$ that is closest to $\mathbf{b}$ is $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$. In other words,

$$\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b}-\mathbf{m}\|_2 = \|\mathbf{b}-\mathbf{P}_{\mathcal{M}}\mathbf{b}\|_2 = dist\,(\mathbf{b},\mathcal{M}). \tag{5.13.13}$$

This is called the ***orthogonal distance*** between $\mathbf{b}$ and $\mathcal{M}$.

Proof. If $\mathbf{p} = \mathbf{P}_{\mathcal{M}}\mathbf{b}$, then $\mathbf{p} - \mathbf{m} \in \mathcal{M}$ for all $\mathbf{m} \in \mathcal{M}$, and

$$\mathbf{b} - \mathbf{p} = (\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{b} \in \mathcal{M}^\perp,$$

so $(\mathbf{p} - \mathbf{m}) \perp (\mathbf{b} - \mathbf{p})$. The Pythagorean theorem says $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ whenever $\mathbf{x} \perp \mathbf{y}$ (recall Exercise 5.4.14), and hence

$$\|\mathbf{b}-\mathbf{m}\|_2^2 = \|\mathbf{b}-\mathbf{p}+\mathbf{p}-\mathbf{m}\|_2^2 = \|\mathbf{b}-\mathbf{p}\|_2^2 + \|\mathbf{p}-\mathbf{m}\|_2^2 \geq \|\mathbf{p}-\mathbf{m}\|_2^2\,.$$

In other words, $\min_{\mathbf{m}\in\mathcal{M}} \|\mathbf{b}-\mathbf{m}\|_2 = \|\mathbf{b}-\mathbf{p}\|_2$. Now argue that there is not another point in $\mathcal{M}$ that is as close to $\mathbf{b}$ as $\mathbf{p}$ is. If $\widehat{\mathbf{m}} \in \mathcal{M}$ such that $\|\mathbf{b}-\widehat{\mathbf{m}}\|_2 = \|\mathbf{b}-\mathbf{p}\|_2$, then by using the Pythagorean theorem again we see

$$\|\mathbf{b}-\widehat{\mathbf{m}}\|_2^2 = \|\mathbf{b}-\mathbf{p}+\mathbf{p}-\widehat{\mathbf{m}}\|_2^2 = \|\mathbf{b}-\mathbf{p}\|_2^2 + \|\mathbf{p}-\widehat{\mathbf{m}}\|_2^2 \implies \|\mathbf{p}-\widehat{\mathbf{m}}\|_2 = 0,$$

and thus $\widehat{\mathbf{m}} = \mathbf{p}$. ∎

Example 5.13.4

To illustrate some of the previous ideas, consider $\Re^{n\times n}$ with the inner product $\langle \mathbf{A}|\mathbf{B}\rangle = trace\left(\mathbf{A}^T\mathbf{B}\right)$. If $\mathcal{S}_n$ is the subspace of $n\times n$ real-symmetric matrices, then each of the following statements is true.

- $\mathcal{S}_n^{\perp}$ = the subspace $\mathcal{K}_n$ of $n\times n$ skew-symmetric matrices.
  - ▷ $\mathcal{S}_n \perp \mathcal{K}_n$ because for all $\mathbf{S}\in\mathcal{S}_n$ and $\mathbf{K}\in\mathcal{K}_n$,

$$\begin{aligned}\langle \mathbf{S}|\mathbf{K}\rangle &= trace\left(\mathbf{S}^T\mathbf{K}\right) = -trace\left(\mathbf{S}\mathbf{K}^T\right) = -trace\left(\mathbf{S}\mathbf{K}^T\right)^T \\ &= -trace\left(\mathbf{K}\mathbf{S}^T\right) = -trace\left(\mathbf{S}^T\mathbf{K}\right) = -\langle \mathbf{S}|\mathbf{K}\rangle \\ \implies &\quad \langle \mathbf{S}|\mathbf{K}\rangle = 0.\end{aligned}$$

  - ▷ $\Re^{n\times n} = \mathcal{S}_n \oplus \mathcal{K}_n$ because every $\mathbf{A}\in\Re^{n\times n}$ can be uniquely expressed as the sum of a symmetric and a skew-symmetric matrix by writing

$$\mathbf{A} = \frac{\mathbf{A}+\mathbf{A}^T}{2} + \frac{\mathbf{A}-\mathbf{A}^T}{2} \qquad \text{(recall (5.9.3) and Exercise 3.2.6).}$$

- The orthogonal projection of $\mathbf{A}\in\Re^{n\times n}$ onto $\mathcal{S}_n$ is $\mathbf{P}(\mathbf{A}) = (\mathbf{A}+\mathbf{A}^T)/2$.
- The closest symmetric matrix to $\mathbf{A}\in\Re^{n\times n}$ is $\mathbf{P}(\mathbf{A}) = (\mathbf{A}+\mathbf{A}^T)/2$.
- The distance from $\mathbf{A}\in\Re^{n\times n}$ to $\mathcal{S}_n$ (the deviation from symmetry) is

$$dist(\mathbf{A},\mathcal{S}_n) = \|\mathbf{A}-\mathbf{P}(\mathbf{A})\|_F = \left\|(\mathbf{A}-\mathbf{A}^T)/2\right\|_F = \sqrt{\frac{trace\left(\mathbf{A}^T\mathbf{A}\right)-trace\left(\mathbf{A}^2\right)}{2}}.$$

Example 5.13.5

Affine Projections. If $\mathbf{v}\neq\mathbf{0}$ is a vector in a space $\mathcal{V}$, and if $\mathcal{M}$ is a subspace of $\mathcal{V}$, then the set of points $\mathcal{A} = \mathbf{v}+\mathcal{M}$ is called an ***affine space*** in $\mathcal{V}$. Strictly speaking, $\mathcal{A}$ is not a subspace (e.g., it doesn't contain $\mathbf{0}$), but, as depicted in Figure 5.13.5, $\mathcal{A}$ is the translate of a subspace—i.e., $\mathcal{A}$ is just a copy of $\mathcal{M}$ that has been translated away from the origin through $\mathbf{v}$. Consequently, notions such as projection onto $\mathcal{A}$ and points closest to $\mathcal{A}$ are analogous to the corresponding concepts for subspaces.

Problem: For $\mathbf{b} \in \mathcal{V}$, determine the point $\mathbf{p}$ in $\mathcal{A} = \mathbf{v} + \mathcal{M}$ that is closest to $\mathbf{b}$. In other words, explain how to project $\mathbf{b}$ orthogonally onto $\mathcal{A}$.

Solution: The trick is to subtract $\mathbf{v}$ from $\mathbf{b}$ as well as from everything in $\mathcal{A}$ to put things back into the context of subspaces where we already know the answers. As illustrated in Figure 5.13.5, this moves $\mathcal{A}$ back down to $\mathcal{M}$, and it translates $\mathbf{v} \to \mathbf{0}$, $\mathbf{b} \to (\mathbf{b} - \mathbf{v})$, and $\mathbf{p} \to (\mathbf{p} - \mathbf{v})$.

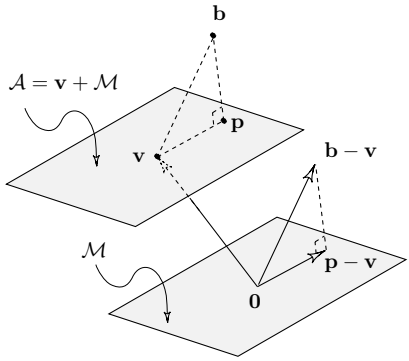


FIGURE 5.13.5

If $\mathbf{p}$ is to be the orthogonal projection of $\mathbf{b}$ onto $\mathcal{A}$, then $\mathbf{p} - \mathbf{v}$ must be the orthogonal projection of $\mathbf{b} - \mathbf{v}$ onto $\mathcal{M}$, so

$$\mathbf{p} - \mathbf{v} = \mathbf{P}_{\mathcal{M}}(\mathbf{b} - \mathbf{v}) \quad \Longrightarrow \quad \mathbf{p} = \mathbf{v} + \mathbf{P}_{\mathcal{M}}(\mathbf{b} - \mathbf{v}), \tag{5.13.14}$$

and thus $\mathbf{p}$ is the point in $\mathcal{A}$ that is closest to $\mathbf{b}$. Applications to the solution of linear systems are developed in Exercises 5.13.17–5.13.22.

We are now in a position to replace the classical calculus-based theory of least squares presented in §4.6 with a more modern vector space development. In addition to being straightforward, the modern geometrical approach puts the entire least squares picture in much sharper focus. Viewing concepts from more than one perspective generally produces deeper understanding, and this is particularly true for the theory of least squares.

Recall from p. 226 that for an inconsistent system $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{b}$, the object of the least squares problem is to find vectors $\mathbf{x}$ that minimize the quantity

$$(\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 . \tag{5.13.15}$$

The classical development in §4.6 relies on calculus to argue that the set of vectors $\mathbf{x}$ that minimize (5.13.15) is exactly the set that solves the (always consistent) system of normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. In the context of the closest point theorem the least squares problem asks for vectors $\mathbf{x}$ such that $\mathbf{A}\mathbf{x}$ is as close

to $\mathbf{b}$ as possible. But $\mathbf{Ax}$ is always a vector in $R(\mathbf{A})$, and the closest point theorem says that the vector in $R(\mathbf{A})$ that is closest to $\mathbf{b}$ is $\mathbf{P}_{R(\mathbf{A})}\mathbf{b}$, the orthogonal projection of $\mathbf{b}$ onto $R(\mathbf{A})$. Figure 5.13.6 illustrates the situation in $\Re^3$.

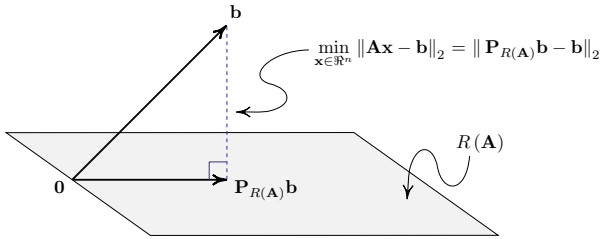


FIGURE 5.13.6

So the least squares problem boils down to finding vectors $\mathbf{x}$ such that

$$\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$$

But this system is equivalent to the system of normal equations because

$$\begin{aligned}
\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b} &\iff \mathbf{P}_{R(\mathbf{A})}\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b} \\
&\iff \mathbf{P}_{R(\mathbf{A})}(\mathbf{Ax}-\mathbf{b}) = \mathbf{0} \\
&\iff (\mathbf{Ax}-\mathbf{b}) \in N\left(\mathbf{P}_{R(\mathbf{A})}\right) = R(\mathbf{A})^{\perp} = N\left(\mathbf{A}^T\right) \\
&\iff \mathbf{A}^T(\mathbf{Ax}-\mathbf{b}) = \mathbf{0} \\
&\iff \mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}.
\end{aligned}$$

Characterizing the set of least squares solutions as the solutions to $\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}$ makes it obvious that $\mathbf{x} = \mathbf{A}^{\dagger}\mathbf{b}$ is a particular least squares solution because (5.13.12) insures $\mathbf{AA}^{\dagger} = \mathbf{P}_{R(\mathbf{A})}$, and thus

$$\mathbf{A}(\mathbf{A}^{\dagger}\mathbf{b}) = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$$

Furthermore, since $\mathbf{A}^{\dagger}\mathbf{b}$ is a particular solution of $\mathbf{Ax} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}$, the general solution—i.e., the set of all least squares solutions—must be the affine space $\mathcal{S} = \mathbf{A}^{\dagger}\mathbf{b} + N(\mathbf{A})$. Finally, the fact that $\mathbf{A}^{\dagger}\mathbf{b}$ is the least squares solution of minimal norm follows from Example 5.13.5 together with

$$R\left(\mathbf{A}^{\dagger}\right) = R\left(\mathbf{A}^T\right) = N(\mathbf{A})^{\perp} \quad \text{(see part (g) of Exercise 5.12.16)}$$

because (5.13.14) insures that the point in $\mathcal{S}$ that is closest to the origin is

$$\mathbf{p} = \mathbf{A}^{\dagger}\mathbf{b} + \mathbf{P}_{N(\mathbf{A})}(\mathbf{0} - \mathbf{A}^{\dagger}\mathbf{b}) = \mathbf{A}^{\dagger}\mathbf{b}.$$

The classical development in §4.6 based on partial differentiation is not easily generalized to cover the case of complex matrices, but the vector space approach given in this example trivially extends to complex matrices by simply replacing $(\star)^T$ by $(\star)^*$.

Below is a summary of some of the major points concerning the theory of least squares.

Least Squares Solutions

Each of the following four statements is equivalent to saying that $\widehat{\mathbf{x}}$ is a least squares solution for a possibly inconsistent linear system $\mathbf{Ax} = \mathbf{b}$.

- $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2 = \min_{\mathbf{x}\in\Re^n} \|\mathbf{Ax} - \mathbf{b}\|_2 .$ (5.13.16)
- $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{P}_{R(\mathbf{A})}\mathbf{b}.$ (5.13.17)
- $\mathbf{A}^T\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}$ $\quad$ ($\mathbf{A}^*\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}^*\mathbf{b}$ when $\mathbf{A} \in \mathcal{C}^{m\times n}$). (5.13.18)
- $\widehat{\mathbf{x}} \in \mathbf{A}^\dagger\mathbf{b} + N(\mathbf{A})$ ($\mathbf{A}^\dagger\mathbf{b}$ is the minimal 2-norm LSS). (5.13.19)

Caution! These are valuable theoretical characterizations, but none is recommended for floating-point computation. Directly solving (5.13.17) or (5.13.18) or explicitly computing $\mathbf{A}^\dagger$ can be inefficient and numerically unstable. Computational issues are discussed in Example 4.5.1 on p. 214; Example 5.5.3 on p. 313; and Example 5.7.3 on p. 346.

The least squares story will not be complete until the following fundamental question is answered: "Why is the method of least squares the best way to make estimates of physical phenomena in the face of uncertainty?" This is the focal point of the next section.

Exercises for section 5.13

5.13.1. Find the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M} = span\{\mathbf{u}\}$, and then determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}^\perp$, where $\mathbf{b} = (\,4 \quad 8\,)^T$ and $\mathbf{u} = (\,3 \quad 1\,)^T$.

5.13.2. Let $\mathbf{A} = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 2 & 0 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

(a) Compute the orthogonal projectors onto each of the four fundamental subspaces associated with $\mathbf{A}$.

(b) Find the point in $N(\mathbf{A})^\perp$ that is closest to $\mathbf{b}$.

5.13.3. For an orthogonal projector $\mathbf{P}$, prove that $\|\mathbf{Px}\|_2 = \|\mathbf{x}\|_2$ if and only if $\mathbf{x} \in R(\mathbf{P})$.

5.13.4. Explain why $\mathbf{A}^T\mathbf{P}_{R(\mathbf{A})} = \mathbf{A}^T$ for all $\mathbf{A} \in \Re^{m\times n}$.

5.13.5. Explain why $\mathbf{P}_{\mathcal{M}} = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^T$ whenever $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for $\mathcal{M} \subseteq \Re^{n\times 1}$.

5.13.6. Explain how to use orthogonal reduction techniques to compute the orthogonal projectors onto each of the four fundamental subspaces of a matrix $\mathbf{A} \in \Re^{m\times n}$.

5.13.7. (a) Describe all 2×2 orthogonal projectors in $\Re^{2\times 2}$.
(b) Describe all 2×2 projectors in $\Re^{2\times 2}$.

5.13.8. The line $\mathcal{L}$ in $\Re^n$ passing through two distinct points $\mathbf{u}$ and $\mathbf{v}$ is $\mathcal{L} = \mathbf{u} + span\,\{\mathbf{u} - \mathbf{v}\}$. If $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{v} \neq \alpha\mathbf{u}$, then $\mathcal{L}$ is a line not passing through the origin—i.e., $\mathcal{L}$ is not a subspace. Sketch a picture in $\Re^2$ or $\Re^3$ to visualize this, and then explain how to project a vector $\mathbf{b}$ orthogonally onto $\mathcal{L}$.

5.13.9. Explain why $\widehat{\mathbf{x}}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$ if and only if $\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2 = \left\|\mathbf{P}_{N(\mathbf{A}^T)}\mathbf{b}\right\|_2$.

5.13.10. Prove that if $\boldsymbol{\varepsilon} = \mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}$, where $\widehat{\mathbf{x}}$ is a least squares solution for $\mathbf{Ax} = \mathbf{b}$, then $\|\boldsymbol{\varepsilon}\|_2^2 = \|\mathbf{b}\|_2^2 - \left\|\mathbf{P}_{R(\mathbf{A})}\mathbf{b}\right\|_2^2$.

5.13.11. Let $\mathcal{M}$ be an r-dimensional subspace of $\Re^n$. We know from (5.4.3) that if $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r\}$ is an orthonormal basis for $\mathcal{M}$, and if $\mathbf{x} \in \mathcal{M}$, then $\mathbf{x}$ is equal to its Fourier expansion with respect to $\mathcal{B}$. That is, $\mathbf{x} = \sum_{i=1}^{r} (\mathbf{u}_i^T\mathbf{x})\mathbf{u}_i$. However, if $\mathbf{x} \notin \mathcal{M}$, then equality is not possible (why?), so the question that arises is, "What does the Fourier expansion on the right-hand side of this expression represent?" Answer this question by showing that the Fourier expansion $\sum_{i=1}^{r} (\mathbf{u}_i^T\mathbf{x})\mathbf{u}_i$ is the point in $\mathcal{M}$ that is closest to $\mathbf{x}$ in the euclidean norm. In other words, show that $\sum_{i=1}^{r} (\mathbf{u}_i^T\mathbf{x})\mathbf{u}_i = \mathbf{P}_{\mathcal{M}}\mathbf{x}$.

5.13.12. Determine the orthogonal projection of $\mathbf{b}$ onto $\mathcal{M}$, where

$$\mathbf{b} = \begin{pmatrix} 5 \\ 2 \\ 5 \\ 3 \end{pmatrix} \quad \text{and} \quad \mathcal{M} = span\left\{ \begin{pmatrix} -3/5 \\ 0 \\ 4/5 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 4/5 \\ 0 \\ 3/5 \\ 0 \end{pmatrix} \right\}.$$

Hint: Is this spanning set in fact an orthonormal basis?

5.13.13. Let $\mathcal{M}$ and $\mathcal{N}$ be subspaces of a vector space $\mathcal{V}$, and consider the associated orthogonal projectors $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$.

(a) Prove that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $\mathcal{M} \perp \mathcal{N}$.

(b) Is it true that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{0}$? Why?

5.13.14. Let $\mathcal{M}$ and $\mathcal{N}$ be subspaces of the same vector space, and let $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ be orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively.

(a) Prove that $R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = R(\mathbf{P}_{\mathcal{M}}) + R(\mathbf{P}_{\mathcal{N}}) = \mathcal{M} + \mathcal{N}$. **Hint:** Use Exercise 4.2.9 along with (4.5.5).

(b) Explain why $\mathcal{M} \perp \mathcal{N}$ if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$.

(c) Explain why $\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}$ is an orthogonal projector if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$, in which case $R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{M} \perp \mathcal{N}$. **Hint:** Recall Exercise 5.9.17.

5.13.15. **Anderson–Duffin Formula.**[59] Prove that if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of the same vector space, then the orthogonal projector onto $\mathcal{M} \cap \mathcal{N}$ is given by $\mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = 2\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}}$. **Hint:** Use (5.13.12) and Exercise 5.13.14 to show $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$. Argue that if $\mathbf{Z} = 2\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$, then $\mathbf{Z} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{Z} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}$.

5.13.16. Given a square matrix $\mathbf{X}$, the ***matrix exponential*** $\mathrm{e}^{\mathbf{X}}$ is defined as

$$\mathrm{e}^{\mathbf{X}} = \mathbf{I} + \mathbf{X} + \frac{\mathbf{X}^2}{2!} + \frac{\mathbf{X}^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{\mathbf{X}^n}{n!}.$$

It can be shown that this series converges for all $\mathbf{X}$, and it is legitimate to differentiate and integrate it term by term to produce the statements $d\mathrm{e}^{\mathbf{A}t}/dt = \mathbf{A}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathbf{A}$ and $\int \mathrm{e}^{\mathbf{A}t}\mathbf{A}\,dt = \mathrm{e}^{\mathbf{A}t}$.

(a) Use the fact that $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t} = \mathbf{0}$ for all $\mathbf{A} \in \Re^{m\times n}$ to show $\mathbf{A}^{\dagger} = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\mathbf{A}^T dt$.

(b) If $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t} = \mathbf{0}$, show $\mathbf{A}^D = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t}\mathbf{A}^k dt$, where $k = index(\mathbf{A})$.[60]

(c) For nonsingular matrices, show that if $\lim_{t\to\infty} \mathrm{e}^{-\mathbf{A}t} = \mathbf{0}$, then $\mathbf{A}^{-1} = \int_0^{\infty} \mathrm{e}^{-\mathbf{A}t} dt$.

---

[59] W. N. Anderson, Jr., and R. J. Duffin discovered this formula for the orthogonal projector onto an intersection in 1969. They called $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}}$ the *parallel sum* of $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ because it is the matrix generalization of the scalar function $r_1 r_2/(r_1 + r_2) = r_1(r_1 + r_2)^{-1} r_2$ that is the resistance of a circuit composed of two resistors r_1 and r_2 connected in parallel. The simple elegance of the Anderson–Duffin formula makes it one of the innumerable little sparkling facets in the jewel that is linear algebra.

[60] A more useful integral representation for $\mathbf{A}^D$ is given in Exercise 7.9.22 (p. 615).

5.13.17. An affine space $\mathbf{v} + \mathcal{M} \subseteq \Re^n$ for which $\dim \mathcal{M} = n - 1$ is called a ***hyperplane.*** For example, a hyperplane in $\Re^2$ is a line (not necessarily through the origin), and a hyperplane in $\Re^3$ is a plane (not necessarily through the origin). The ith equation $\mathbf{A}_{i*}\mathbf{x} = b_i$ in a linear system $\mathbf{A}_{m\times n}\mathbf{x} = \mathbf{b}$ is a hyperplane in $\Re^n$, so the solutions of $\mathbf{Ax} = \mathbf{b}$ occur at the intersection of the m hyperplanes defined by the rows of $\mathbf{A}$.

(a) Prove that for a given scalar β and a nonzero vector $\mathbf{u} \in \Re^n$, the set $\mathcal{H} = \{\mathbf{x} \,|\, \mathbf{u}^T\mathbf{x} = \beta\}$ is a hyperplane in $\Re^n$.

(b) Explain why the orthogonal projection of $\mathbf{b} \in \Re^n$ onto $\mathcal{H}$ is $\mathbf{p} = \mathbf{b} - \left(\mathbf{u}^T\mathbf{b} - \beta/\mathbf{u}^T\mathbf{u}\right)\mathbf{u}$.

5.13.18. For $\mathbf{u}, \mathbf{w} \in \Re^n$ such that $\mathbf{u}^T\mathbf{w} \neq 0$, let $\mathcal{M} = \mathbf{u}^\perp$ and $\mathcal{W} = span\,\{\mathbf{w}\}$.

(a) Explain why $\Re^n = \mathcal{M} \oplus \mathcal{W}$.

(b) For $\mathbf{b} \in \Re^{n\times 1}$, explain why the *oblique* projection of $\mathbf{b}$ onto $\mathcal{M}$ along $\mathcal{W}$ is given by $\mathbf{p} = \mathbf{b} - \mathbf{u}^T\mathbf{b}/\mathbf{u}^T\mathbf{w}\mathbf{w}$.

(c) For a given scalar β, let $\mathcal{H}$ be the hyperplane in $\Re^n$ defined by $\mathcal{H} = \{\mathbf{x} \,|\, \mathbf{u}^T\mathbf{x} = \beta\}$—see Exercise 5.13.17. Explain why the *oblique* projection of $\mathbf{b}$ onto $\mathcal{H}$ along $\mathcal{W}$ should be given by $\mathbf{p} = \mathbf{b} - \left(\mathbf{u}^T\mathbf{b} - \beta/\mathbf{u}^T\mathbf{w}\right)\mathbf{w}$.

5.13.19. Kaczmarz's [61] Projection Method. The solution of a nonsingular system

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

is the intersection of the two hyperplanes (lines in this case) defined by

$$\mathcal{H}_1 = \{(x_1, x_2) \,|\, a_{11}x_1 + a_{12}x_2 = b_1\}\,,\ \mathcal{H}_2 = \{(x_1, x_2) \,|\, a_{21}x_1 + a_{22}x_2 = b_2\}.$$

It's visually evident that by starting with an arbitrary point $\mathbf{p}_0$ and alternately projecting orthogonally onto $\mathcal{H}_1$ and $\mathcal{H}_2$ as depicted in Figure 5.13.7, the resulting sequence of projections $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4, \ldots\}$ converges to $\mathcal{H}_1 \cap \mathcal{H}_2$, the solution of $\mathbf{Ax} = \mathbf{b}$.

[61] Although this idea has probably occurred to many people down through the ages, credit is usually given to Stefan Kaczmarz, who published his results in 1937. Kaczmarz was among a school of bright young Polish mathematicians who were beginning to flower in the first part of the twentieth century. Tragically, this group was decimated by Hitler's invasion of Poland, and Kaczmarz himself was killed in military action while trying to defend his country.

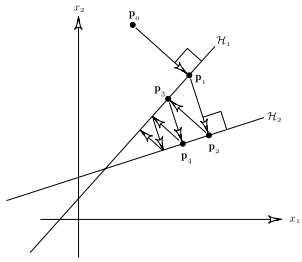


FIGURE 5.13.7

This idea can be generalized by using Exercise 5.13.17. For a consistent system $\mathbf{A}_{n\times r}\mathbf{x} = \mathbf{b}$ with $rank\,(\mathbf{A}) = r$, scale the rows so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, and let $\mathcal{H}_i = \{\mathbf{x} \,|\, \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the ith equation. Begin with an arbitrary vector $\mathbf{p}_0 \in \Re^{r\times 1}$, and successively perform orthogonal projections onto each hyperplane to generate the following sequence:

$$\begin{array}{ll}
\mathbf{p}_1 = \mathbf{p}_0 - (\mathbf{A}_{1*}\mathbf{p}_0 - b_1)\,(\mathbf{A}_{1*})^T & (\text{project } \mathbf{p}_0 \text{ onto } \mathcal{H}_1\,), \\
\mathbf{p}_2 = \mathbf{p}_1 - (\mathbf{A}_{2*}\mathbf{p}_1 - b_2)\,(\mathbf{A}_{2*})^T & (\text{project } \mathbf{p}_1 \text{ onto } \mathcal{H}_2\,), \\
\vdots & \vdots \\
\mathbf{p}_n = \mathbf{p}_{n-1} - (\mathbf{A}_{n*}\mathbf{p}_{n-1} - b_n)\,(\mathbf{A}_{n*})^T & (\text{project } \mathbf{p}_{n-1} \text{ onto } \mathcal{H}_n\,).
\end{array}$$

When all n hyperplanes have been used, continue by repeating the process. For example, on the second pass project $\mathbf{p}_n$ onto $\mathcal{H}_1$; then project $\mathbf{p}_{n+1}$ onto $\mathcal{H}_2$, etc. For an arbitrary $\mathbf{p}_0$, the entire Kaczmarz sequence is generated by executing the following double loop:

$$\begin{array}{l}
\text{For } k = 0, 1, 2, 3, \ldots \\
\qquad \text{For } i = 1, 2, \ldots, n \\
\qquad\qquad \mathbf{p}_{kn+i} = \mathbf{p}_{kn+i-1} - (\mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i)\,(\mathbf{A}_{i*})^T
\end{array}$$

Prove that the Kaczmarz sequence converges to the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ by showing $\|\mathbf{p}_{kn+i} - \mathbf{x}\|_2^2 = \|\mathbf{p}_{kn+i-1} - \mathbf{x}\|_2^2 - (\mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i)^2$.

5.13.20. Oblique Projection Method. Assume that a nonsingular system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$ has been row scaled so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, and let $\mathcal{H}_i = \{\mathbf{x} \,|\, \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the ith equation—see Exercise 5.13.17. In theory, the system can be solved by making $n-1$ oblique projections of the type described in Exercise 5.13.18 because if an arbitrary point $\mathbf{p}_1$ in $\mathcal{H}_1$ is projected obliquely onto $\mathcal{H}_2$ along $\mathcal{H}_1$ to produce $\mathbf{p}_2$, then $\mathbf{p}_2$ is in $\mathcal{H}_1 \cap \mathcal{H}_2$. If $\mathbf{p}_2$ is projected onto $\mathcal{H}_3$ along $\mathcal{H}_1 \cap \mathcal{H}_2$ to produce $\mathbf{p}_3$, then $\mathbf{p}_3 \in \mathcal{H}_1 \cap \mathcal{H}_2 \cap \mathcal{H}_3$, and so forth until $\mathbf{p}_n \in \cap_{i=1}^n \mathcal{H}_i$. This is similar to Kaczmarz's method given in Exercise 5.13.19, but here we are projecting obliquely instead of orthogonally. However, projecting $\mathbf{p}_k$ onto $\mathcal{H}_{k+1}$ along $\cap_{i=1}^k \mathcal{H}_i$ is difficult because

$\cap_{i=1}^{k}\mathcal{H}_i$ is generally unknown. This problem is overcome by modifying the procedure as follows—use Figure 5.13.8 with $n = 3$ as a guide.

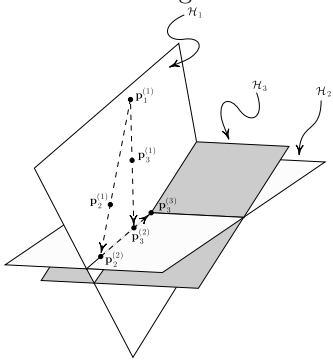


FIGURE 5.13.8

Step 0. Begin with any set $\{\mathbf{p}_1^{(1)}, \mathbf{p}_2^{(1)}, \ldots, \mathbf{p}_n^{(1)}\} \subset \mathcal{H}_1$ such that $\{(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}), (\mathbf{p}_1^{(1)} - \mathbf{p}_3^{(1)}), \ldots, (\mathbf{p}_1^{(1)} - \mathbf{p}_n^{(1)})\}$ is linearly independent and $\mathbf{A}_{2*}(\mathbf{p}_1^{(1)} - \mathbf{p}_k^{(1)}) \neq 0$ for $k = 2, 3, \ldots, n$.

Step 1. In turn, project $\mathbf{p}_1^{(1)}$ onto $\mathcal{H}_2$ through $\mathbf{p}_2^{(1)}, \mathbf{p}_3^{(1)}, \ldots, \mathbf{p}_n^{(1)}$ to produce $\{\mathbf{p}_2^{(2)}, \mathbf{p}_3^{(2)}, \ldots, \mathbf{p}_n^{(2)}\} \subset \mathcal{H}_1 \cap \mathcal{H}_2$ (see Figure 5.13.8).

Step 2. Project $\mathbf{p}_2^{(2)}$ onto $\mathcal{H}_3$ through $\mathbf{p}_3^{(2)}, \mathbf{p}_4^{(2)}, \ldots, \mathbf{p}_n^{(2)}$ to produce $\{\mathbf{p}_3^{(3)}, \mathbf{p}_4^{(3)}, \ldots, \mathbf{p}_n^{(3)}\} \subset \mathcal{H}_1 \cap \mathcal{H}_2 \cap \mathcal{H}_3$. And so the process continues.

Step n−1. Project $\mathbf{p}_{n-1}^{(n-1)}$ through $\mathbf{p}_n^{(n-1)}$ to produce $\mathbf{p}_n^{(n)} \in \cap_{i=1}^{n}\mathcal{H}_i$. Of course, $\mathbf{x} = \mathbf{p}_n^{(n)}$ is the solution of the system.

For any initial set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathcal{H}_1$ satisfying the properties described in Step 0, explain why the following algorithm performs the computations described in Steps $1, 2, \ldots, n-1$.

For $i = 2$ to n

For $j = i$ to n

$$\mathbf{x}_j \leftarrow \mathbf{x}_j - \frac{(\mathbf{A}_{i*}\mathbf{x}_{i-1} - b_i)}{\mathbf{A}_{i*}(\mathbf{x}_{i-1} - \mathbf{x}_j)}(\mathbf{x}_{i-1} - \mathbf{x}_j)$$

$\mathbf{x} \leftarrow \mathbf{x}_n$ (the solution of the system)

5.13.21. Let $\mathcal{M}$ be a subspace of $\Re^n$, and let $\mathbf{R} = \mathbf{I} - 2\mathbf{P}_{\mathcal{M}}$. Prove that the orthogonal distance between any point $\mathbf{x} \in \Re^n$ and $\mathcal{M}^{\perp}$ is the same as the orthogonal distance between $\mathbf{Rx}$ and $\mathcal{M}^{\perp}$. In other words, prove that $\mathbf{R}$ reflects everything in $\Re^n$ about $\mathcal{M}^{\perp}$. Naturally, $\mathbf{R}$ is called the ***reflector*** about $\mathcal{M}^{\perp}$. The *elementary* reflectors $\mathbf{I} - 2\mathbf{u}\mathbf{u}^T/\mathbf{u}^T\mathbf{u}$ discussed on p. 324 are special cases—go back and look at Figure 5.6.2.

5.13.22. Cimmino's Reflection Method. In 1938 the Italian mathematician Gianfranco Cimmino used the following elementary observation to construct an iterative algorithm for solving linear systems. For a 2×2 system $\mathbf{Ax} = \mathbf{b}$, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the two lines (hyperplanes) defined by the two equations. For an arbitrary guess $\mathbf{r}_0$, let $\mathbf{r}_1$ be the reflection of $\mathbf{r}_0$ about the line $\mathcal{H}_1$, and let $\mathbf{r}_2$ be the reflection of $\mathbf{r}_0$ about the line $\mathcal{H}_2$. As illustrated in Figure 5.13.9, the three points $\mathbf{r}_0$, $\mathbf{r}_1$, and $\mathbf{r}_2$ lie on a circle whose center is $\mathcal{H}_1 \cap \mathcal{H}_2$ (the solution of the system).

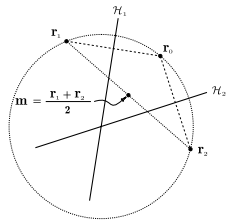


FIGURE 5.13.9

The mean value $\mathbf{m} = (\mathbf{r}_1 + \mathbf{r}_2)/2$ is strictly inside the circle, so $\mathbf{m}$ is a better approximation to the solution than $\mathbf{r}_0$. It's visually evident that iteration produces a sequence that converges to the solution of $\mathbf{Ax} = \mathbf{b}$. Prove this in general by using the following blueprint.

(a) For a scalar β and a vector $\mathbf{u} \in \Re^n$ such that $\|\mathbf{u}\|_2 = 1$, consider the hyperplane $\mathcal{H} = \{\mathbf{x} \,|\, \mathbf{u}^T\mathbf{x} = \beta\}$ (Exercise 5.13.17). Use (5.6.8) to show that the reflection of a vector $\mathbf{b}$ about $\mathcal{H}$ is $\mathbf{r} = \mathbf{b} - 2(\mathbf{u}^T\mathbf{b} - \beta)\mathbf{u}$.

(b) For a system $\mathbf{Ax} = \mathbf{b}$ in which the rows of $\mathbf{A} \in \Re^{n\times r}$ have been scaled so that $\|\mathbf{A}_{i*}\|_2 = 1$ for each i, let $\mathcal{H}_i = \{\mathbf{x} \,|\, \mathbf{A}_{i*}\mathbf{x} = b_i\}$ be the hyperplane defined by the ith equation. If $\mathbf{r}_0 \in \Re^{r\times 1}$ is an arbitrary vector, and if $\mathbf{r}_i$ is the reflection of $\mathbf{r}_0$ about $\mathcal{H}_i$, explain why the mean value of the reflections $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ is $\mathbf{m} = \mathbf{r}_0 - (2/n)\mathbf{A}^T\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = \mathbf{Ar}_0 - \mathbf{b}$.

(c) Iterating part (b) produces $\mathbf{m}_k = \mathbf{m}_{k-1} - (2/n)\mathbf{A}^T\boldsymbol{\varepsilon}_{k-1}$, where $\boldsymbol{\varepsilon}_{k-1} = \mathbf{Am}_{k-1} - \mathbf{b}$. Show that if $\mathbf{A}$ is nonsingular, and if $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, then $\mathbf{x} - \mathbf{m}_k = \left(\mathbf{I} - (2/n)\mathbf{A}^T\mathbf{A}\right)^k(\mathbf{x} - \mathbf{m}_0)$. **Note:** It can be proven that $\left(\mathbf{I} - (2/n)\mathbf{A}^T\mathbf{A}\right)^k \to \mathbf{0}$ as $k \to \infty$, so $\mathbf{m}_k \to \mathbf{x}$ for all $\mathbf{m}_0$. In fact, $\mathbf{m}_k$ converges even if $\mathbf{A}$ is rank deficient—if consistent, it converges to a solution, and, if inconsistent, the limit is a least squares solution. Cimmino's method also works with weighted means. If $\mathbf{W} = \operatorname{diag}(w_1, w_2, \ldots, w_n)$, where $w_i > 0$ and $\sum w_i = 1$, then $\mathbf{m}_k = \mathbf{m}_{k-1} - \omega\mathbf{A}^T\mathbf{W}\boldsymbol{\varepsilon}_{k-1}$ is a convergent sequence in which $0 < \omega < 2$ is a "relaxation parameter" that can be adjusted to alter the rate of convergence.

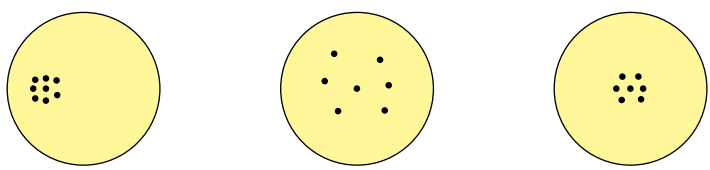
5.14 WHY LEAST SQUARES?

Drawing inferences about natural phenomena based upon physical observations and estimating characteristics of large populations by examining small samples are fundamental concerns of applied science. Numerical characteristics of a phenomenon or population are often called *parameters,* and the goal is to design functions or rules called *estimators* that use observations or samples to estimate parameters of interest. For example, the mean height h of all people is a parameter of the world's population, and one way of estimating h is to observe the mean height of a sample of k people. In other words, if h_i is the height of the ith person in a sample, the function $\hat{h}$ defined by

$$\hat{h}(h_1, h_2, \ldots, h_k) = \frac{1}{k}\left(\sum_{i=1}^{k} h_i\right)$$

is an estimator for h. Moreover, $\hat{h}$ is a *linear estimator* because $\hat{h}$ is a linear function of the observations.

Good estimators should possess at least two properties—they should be *unbiased* and they should have *minimal variance.* For example, consider estimating the center of a circle drawn on a wall by asking Larry, Moe, and Curly to each throw one dart at the circle. To decide which estimator is best, we need to know more about each thrower's style. While being able to throw a tight pattern, it is known that Larry tends to have a left-hand bias in his style. Moe doesn't suffer from a bias, but he tends to throw a rather large pattern. However, Curly can throw a tight pattern without a bias. Typical patterns are shown below.



Larry Moe Curly

Although Larry has a small variance, he is an unacceptable estimator because he is biased in the sense that his average is significantly different than the center. Moe and Curly are each unbiased estimators because they have an average that is the center, but Curly is clearly the preferred estimator because his variance is much smaller than Moe's. In other words, Curly is the unbiased estimator of minimal variance.

To make these ideas more formal, let's adopt the following standard notation and terminology from elementary probability theory concerning random variables X and Y.

- $E[X] = \mu_X$ denotes the ***mean*** (or expected value) of X.
- $\text{Var}[X] = E\left[(X - \mu_X)^2\right] = E[X^2] - \mu_X^2$ is the ***variance*** of X.
- $\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$ is the ***covariance*** of X and Y.

Minimum Variance Unbiased Estimators

An estimator $\hat{\theta}$ (consider as a random variable) for a parameter θ is said to be ***unbiased*** when $E[\hat{\theta}] = \theta$, and $\hat{\theta}$ is called a ***minimum variance unbiased estimator*** for θ whenever $\text{Var}[\hat{\theta}] \leq \text{Var}[\hat{\phi}]$ for all unbiased estimators $\hat{\phi}$ of θ.

These ideas make it possible to precisely articulate why the method of least squares is the best way to fit observed data. Let Y be a variable that is known (or assumed) to be linearly related to other variables $X_1, X_2, \ldots, X_n$ according to the equation[62]

$$Y = \beta_1 X_1 + \cdots + \beta_n X_n, \qquad (5.14.1),$$

where the β_i's are unknown constants (parameters). Suppose that the values assumed by the X_i's are not subject to error or variation and can be exactly observed or specified, but, due perhaps to measurement error, the values of Y cannot be exactly observed. Instead, we observe

$$y = Y + \varepsilon = \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon, \qquad (5.14.2)$$

where ε is a random variable accounting for the measurement error. For example, consider the problem of determining the velocity v of a moving object by measuring the distance D it has traveled at various points in time T by using the linear relation $D = vT$. Time can be prescribed at exact values such as $T_1 = 1$ second, $T_2 = 2$ seconds, etc., but observing the distance traveled at the prescribed values of T will almost certainly involve small measurement errors so that in reality the observed distances satisfy $d = D + \varepsilon = vT + \varepsilon$. Now consider the general problem of determining the parameters β_k in (5.14.1) by observing (or measuring) values of Y at m different points $\mathbf{X}_{i*} = (x_{i1}, x_{i2}, \ldots, x_{in}) \in \Re^n$, where x_{ij} is the value of X_j to be used when making the ith observation. If y_i denotes the random variable that represents the outcome of the ith observation of Y, then according to (5.14.2),

$$y_i = \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \varepsilon_i, \quad i = 1, 2, \ldots, m, \qquad (5.14.3)$$

[62] Equation (5.14.1) is called a ***no-intercept model,*** whereas the slightly more general equation $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ is known as an ***intercept model.*** Since the analysis for an intercept model is not significantly different from the analysis of the no-intercept case, we deal only with the no-intercept case and leave the intercept model for the reader to develop.

where ε_i is a random variable accounting for the ith observation (or measurement) error.[63] It is generally valid to assume that observation errors are not correlated with each other but have a common variance (not necessarily known) and a zero mean. In other words, we assume that

$$E[\varepsilon_i] = 0 \text{ for each } i \quad \text{and} \quad \operatorname{Cov}[\varepsilon_i, \varepsilon_j] = \begin{cases} \sigma^2 & \text{when } i = j, \\ 0 & \text{when } i \neq j. \end{cases}$$

If $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$, then the equations in (5.14.3) can be written as $\mathbf{y} = \mathbf{X}_{m\times n}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. In practice, the points $\mathbf{X}_{i*}$ at which observations y_i are made can almost always be selected to insure that $rank\,(\mathbf{X}_{m\times n}) = n$, so the complete statement of the ***standard linear model*** is

$$\mathbf{y} = \mathbf{X}_{m\times n}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{such that} \quad \begin{cases} rank\,(\mathbf{X}) = n, \\ E[\boldsymbol{\varepsilon}] = \mathbf{0}, \\ \operatorname{Cov}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}, \end{cases} \tag{5.14.4}$$

where we have adopted the conventions

$$E[\boldsymbol{\varepsilon}] = \begin{pmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \vdots \\ E[\varepsilon_m] \end{pmatrix} \text{ and } \operatorname{Cov}[\boldsymbol{\varepsilon}] = \begin{pmatrix} \operatorname{Cov}[\varepsilon_1, \varepsilon_1] & \operatorname{Cov}[\varepsilon_1, \varepsilon_2] & \cdots & \operatorname{Cov}[\varepsilon_1, \varepsilon_m] \\ \operatorname{Cov}[\varepsilon_2, \varepsilon_1] & \operatorname{Cov}[\varepsilon_2, \varepsilon_2] & \cdots & \operatorname{Cov}[\varepsilon_2, \varepsilon_m] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}[\varepsilon_m, \varepsilon_1] & \operatorname{Cov}[\varepsilon_m, \varepsilon_2] & \cdots & \operatorname{Cov}[\varepsilon_m, \varepsilon_m] \end{pmatrix}.$$

The problem is to determine the best (minimum variance) linear (linear function of the y_i 's) unbiased estimators for the components of $\boldsymbol{\beta}$. Gauss realized in 1821 that this is precisely what the least squares solution provides.

Gauss–Markov Theorem

For the standard linear model (5.14.4), the minimum variance linear unbiased estimator for β_i is given by the ith component $\hat{\beta}_i$ in the vector $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$. In other words, the best linear unbiased estimator for $\boldsymbol{\beta}$ is the least squares solution of $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}$.

[63] In addition to observation and measurement errors, other errors such as modeling errors or those induced by imposing simplifying assumptions produce the same kind of equation—recall the discussion of ice cream on p. 228.

Proof. It is clear that $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$ is a linear estimator of $\boldsymbol{\beta}$ because each component $\hat{\beta}_i = \sum_k [\mathbf{X}^\dagger]_{ik}\, y_k$ is a linear function of the observations. The fact that $\hat{\boldsymbol{\beta}}$ is unbiased follows by using the linear nature of expected value to write

$$E[\mathbf{y}] = E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = E[\mathbf{X}\boldsymbol{\beta}] + E[\boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta},$$

so that

$$E\big[\hat{\boldsymbol{\beta}}\big] = E\big[\mathbf{X}^\dagger \mathbf{y}\big] = \mathbf{X}^\dagger E[\mathbf{y}] = \mathbf{X}^\dagger \mathbf{X}\boldsymbol{\beta} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

To argue that $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger \mathbf{y}$ has minimal variance among all linear unbiased estimators for $\boldsymbol{\beta}$, let $\boldsymbol{\beta}^*$ be an arbitrary linear unbiased estimator for $\boldsymbol{\beta}$. Linearity of $\boldsymbol{\beta}^*$ implies the existence of a matrix $\mathbf{L}_{n\times m}$ such that $\boldsymbol{\beta}^* = \mathbf{L}\mathbf{y}$, and unbiasedness insures $\boldsymbol{\beta} = E[\boldsymbol{\beta}^*] = E[\mathbf{L}\mathbf{y}] = \mathbf{L}E[\mathbf{y}] = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$. We want $\boldsymbol{\beta} = \mathbf{L}\mathbf{X}\boldsymbol{\beta}$ to hold irrespective of the values of the components in $\boldsymbol{\beta}$, so it must be the case that $\mathbf{L}\mathbf{X} = \mathbf{I}_n$ (recall Exercise 3.5.5). For $i \neq j$ we have

$$0 = \mathrm{Cov}[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i\varepsilon_j] - \mu_{\varepsilon_i}\mu_{\varepsilon_j} \quad \Longrightarrow \quad E[\varepsilon_i\varepsilon_j] = E[\varepsilon_i]E[\varepsilon_j] = 0,$$

so that

$$\mathrm{Cov}[y_i, y_j] = \begin{cases} E[(y_i - \mu_{y_i})^2] = E[\varepsilon_i^2] = \mathrm{Var}[\varepsilon_i] = \sigma^2 & \text{when } i = j, \\ E[(y_i - \mu_{y_i})(y_j - \mu_{y_j})] = E[\varepsilon_i\varepsilon_j] = 0 & \text{when } i \neq j. \end{cases} \tag{5.14.5}$$

This together with the fact that $\mathrm{Var}[aW + bZ] = a^2\mathrm{Var}[W] + b^2\mathrm{Var}[Z]$ whenever $\mathrm{Cov}[W, Z] = 0$ allows us to write

$$\mathrm{Var}[\beta_i^*] = \mathrm{Var}[\mathbf{L}_{i*}\mathbf{y}] = \mathrm{Var}\left[\sum_{k=1}^m l_{ik}y_k\right] = \sigma^2\sum_{k=1}^m l_{ik}^2 = \sigma^2\left\|\mathbf{L}_{i*}\right\|_2^2.$$

Since $\mathbf{L}\mathbf{X} = \mathbf{I}$, it follows that $\mathrm{Var}[\beta_i^*]$ is minimal if and only if $\mathbf{L}_{i*}$ is the minimum norm solution of the system $\mathbf{z}^T\mathbf{X} = \mathbf{e}_i^T$. We know from (5.12.17) that the (unique) minimum norm solution is given by $\mathbf{z}^T = \mathbf{e}_i^T\mathbf{X}^\dagger = \mathbf{X}_{i*}^\dagger$, so $\mathrm{Var}[\beta_i^*]$ is minimal if and only if $\mathbf{L}_{i*} = \mathbf{X}_{i*}^\dagger$. Since this holds for $i = 1, 2, \ldots, m$, it follows that $\mathbf{L} = \mathbf{X}^\dagger$. In other words, the components of $\hat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{y}$ are the (unique) minimal variance linear unbiased estimators for the parameters in $\boldsymbol{\beta}$. ∎

Exercises for section 5.14

5.14.1. For a matrix $\mathbf{Z}_{m\times n} = [z_{ij}]$, of random variables, $E[\mathbf{Z}]$ is defined to be the $m \times n$ matrix whose (i, j)-entry is $E[z_{ij}]$. Consider the standard linear model described in (5.14.4), and let $\hat{\mathbf{e}}$ denote the vector of random variables defined by $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ in which $\hat{\boldsymbol{\beta}} = \big(\mathbf{X}^T\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^\dagger\mathbf{y}$. Demonstrate that

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{m - n}$$

is an unbiased estimator for σ^2. **Hint:** $\mathbf{d}^T\mathbf{c} = trace(\mathbf{c}\mathbf{d}^T)$ for column vectors $\mathbf{c}$ and $\mathbf{d}$, and, by virtue of Exercise 5.9.13,

$$trace\,\big(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger\big) = m - trace\,\big(\mathbf{X}\mathbf{X}^\dagger\big) = m - rank\,\big(\mathbf{X}\mathbf{X}^\dagger\big) = m - n.$$

5.15 ANGLES BETWEEN SUBSPACES

Consider the problem of somehow gauging the separation between a pair of nontrivial but otherwise general subspaces $\mathcal{M}$ and $\mathcal{N}$ of $\Re^n$. Perhaps the first thing that comes to mind is to measure the angle between them. But defining the "angle" between subspaces in $\Re^n$ is not as straightforward as the visual geometry of $\Re^2$ or $\Re^3$ might suggest. There is just too much "wiggle room" in higher dimensions to make any one definition completely satisfying, and the "correct" definition usually varies with the specific application under consideration.

Before exploring general angles, recall what has already been said about some special cases beginning with the angle between a pair of one-dimensional subspaces. If $\mathcal{M}$ and $\mathcal{N}$ are spanned by vectors $\mathbf{u}$ and $\mathbf{v}$, respectively, and if $\|\mathbf{u}\| = 1 = \|\mathbf{v}\|$, then the angle between $\mathcal{M}$ and $\mathcal{N}$ is defined by the expression $\cos\theta = \mathbf{v}^T\mathbf{u}$ (p. 295). This idea was carried one step further on p. 389 to define the angle between two *complementary* subspaces, and an intuitive connection to norms of projectors was presented. These intuitive ideas are now made rigorous.

Minimal Angle

The ***minimal angle*** between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ is defined to be the number $0 \le \theta_{min} \le \pi/2$ for which

$$\cos\theta_{min} = \max_{\substack{\mathbf{u}\in\mathcal{M},\,\mathbf{v}\in\mathcal{N}\\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u}. \tag{5.15.1}$$

- If $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, then

$$\cos\theta_{min} = \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2\,. \tag{5.15.2}$$

- If $\mathcal{M}$ and $\mathcal{N}$ are *complementary* subspaces, and if $\mathbf{P}_{\mathcal{M}\mathcal{N}}$ is the *oblique* projector onto $\mathcal{M}$ along $\mathcal{N}$, then

$$\sin\theta_{min} = \frac{1}{\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2}. \tag{5.15.3}$$

- $\mathcal{M}$ and $\mathcal{N}$ are complementary subspaces if and only if $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is invertible, and in this case

$$\sin\theta_{min} = \frac{1}{\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\|_2}. \tag{5.15.4}$$

Proof of (5.15.2). If $f : \mathcal{V} \to \Re$ is a function defined on a space $\mathcal{V}$ such that $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ for all scalars $\alpha \ge 0$, then

$$\max_{\|\mathbf{x}\|=1} f(\mathbf{x}) = \max_{\|\mathbf{x}\|\le 1} f(\mathbf{x}) \quad \text{(see Exercise 5.15.8).} \tag{5.15.5}$$

This together with (5.2.9) and the fact that $\mathbf{P}_{\mathcal{M}}\mathbf{x} \in \mathcal{M}$ and $\mathbf{P}_{\mathcal{N}}\mathbf{y} \in \mathcal{N}$ means

$$\begin{aligned}\cos\theta_{min} &= \max_{\substack{\mathbf{u}\in\mathcal{M},\,\mathbf{v}\in\mathcal{N}\\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = \max_{\substack{\mathbf{u}\in\mathcal{M},\,\mathbf{v}\in\mathcal{N}\\ \|\mathbf{u}\|_2\le 1,\,\|\mathbf{v}\|_2\le 1}} \mathbf{v}^T\mathbf{u}\\ &= \max_{\|\mathbf{x}\|_2\le 1,\,\|\mathbf{y}\|_2\le 1} \mathbf{y}^T\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\mathbf{x} = \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2 . \quad \blacksquare\end{aligned}$$

Proof of (5.15.3). Let $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ be orthogonal matrices in which the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$ constitute orthonormal bases for $\mathcal{M}$ and $\mathcal{M}^{\perp}$, respectively, and $\mathbf{V}_1$ and $\mathbf{V}_2$ are orthonormal bases for $\mathcal{N}^{\perp}$ and $\mathcal{N}$, respectively, so that $\mathbf{U}_i^T\mathbf{U}_i = \mathbf{I}$ and $\mathbf{V}_i^T\mathbf{V}_i = \mathbf{I}$ for $i = 1, 2$, and

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}_1\mathbf{U}_1^T,\ \ \mathbf{I} - \mathbf{P}_{\mathcal{M}} = \mathbf{U}_2\mathbf{U}_2^T,\ \ \mathbf{P}_{\mathcal{N}} = \mathbf{V}_2\mathbf{V}_2^T,\ \ \mathbf{I} - \mathbf{P}_{\mathcal{N}} = \mathbf{V}_1\mathbf{V}_1^T.$$

As discussed on p. 407, there is a nonsingular matrix $\mathbf{C}$ such that

$$\mathbf{P}_{\mathcal{M}\mathcal{N}} = \mathbf{U}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T = \mathbf{U}_1\mathbf{C}\mathbf{V}_1^T. \tag{5.15.6}$$

Notice that $\mathbf{P}_{\mathcal{M}\mathcal{N}}^2 = \mathbf{P}_{\mathcal{M}\mathcal{N}}$ implies $\mathbf{C} = \mathbf{C}\mathbf{V}_1^T\mathbf{U}_1\mathbf{C}$, which in turn insures $\mathbf{C}^{-1} = \mathbf{V}_1^T\mathbf{U}_1$. Recall that $\|\mathbf{X}\mathbf{A}\mathbf{Y}\|_2 = \|\mathbf{A}\|_2$ whenever $\mathbf{X}$ has orthonormal columns and $\mathbf{Y}$ has orthonormal rows (Exercise 5.6.9). Consequently,

$$\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2 = \|\mathbf{C}\|_2 = \frac{1}{\min\limits_{\|\mathbf{x}\|_2=1}\big\|\mathbf{C}^{-1}\mathbf{x}\big\|_2} = \frac{1}{\min\limits_{\|\mathbf{x}\|_2=1}\big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2} \quad \text{(recall (5.2.6))}.$$

Combining this with (5.15.2) produces (5.15.3) by writing

$$\begin{aligned}\sin^2\theta_{min} &= 1 - \cos^2\theta_{min} = 1 - \|\mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}\|_2^2 = 1 - \big\|\mathbf{V}_2\mathbf{V}_2^T\mathbf{U}_1\mathbf{U}_1^T\big\|_2^2\\ &= 1 - \big\|(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\big\|_2^2 = 1 - \max_{\|\mathbf{x}\|_2=1}\big\|(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\mathbf{x}\big\|_2^2\\ &= 1 - \max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{U}_1^T(\mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_1\mathbf{x} = 1 - \max_{\|\mathbf{x}\|_2=1}\Big(1 - \big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2^2\Big)\\ &= 1 - \Big(1 - \min_{\|\mathbf{x}\|_2=1}\big\|\mathbf{V}_1^T\mathbf{U}_1\mathbf{x}\big\|_2^2\Big) = \frac{1}{\|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2^2}. \quad \blacksquare\end{aligned}$$

Proof of (5.15.4). Observe that

$$\begin{aligned}\mathbf{U}^T(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})\mathbf{V} &= \begin{pmatrix}\mathbf{U}_1^T\\ \hline \mathbf{U}_2^T\end{pmatrix}(\mathbf{U}_1\mathbf{U}_1^T - \mathbf{V}_2\mathbf{V}_2^T)\big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)\\ &= \begin{pmatrix}\mathbf{U}_1^T\mathbf{V}_1 & \mathbf{0}\\ \mathbf{0} & -\mathbf{U}_2^T\mathbf{V}_2\end{pmatrix},\end{aligned} \tag{5.15.7}$$

where $\mathbf{U}_1^T\mathbf{V}_1 = (\mathbf{C}^{-1})^T$ is nonsingular. To see that $\mathbf{U}_2^T\mathbf{V}_2$ is also nonsingular, suppose $\dim\mathcal{M} = r$ so that $\dim\mathcal{N} = n-r$ and $\mathbf{U}_2^T\mathbf{V}_2$ is $n-r \times n-r$. Use the formula for the rank of a product (4.5.1) to write

$$rank\left(\mathbf{U}_2^T\mathbf{V}_2\right) = rank\left(\mathbf{U}_2^T\right) - \dim N\left(\mathbf{U}_2^T\right) \cap R\left(\mathbf{V}_2\right) = n-r-\dim\mathcal{M}\cap\mathcal{N} = n-r.$$

It now follows from (5.15.7) that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular, and

$$\mathbf{V}^T(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\mathbf{U} = \begin{pmatrix} (\mathbf{U}_1^T\mathbf{V}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & -(\mathbf{U}_2^T\mathbf{V}_2)^{-1} \end{pmatrix}.$$

(Showing that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular implies $\mathcal{M} \oplus \mathcal{N} = \Re^n$ is Exercise 5.15.6.) Formula (5.2.12) on p. 283 for the 2-norm of a block-diagonal matrix can now be applied to yield

$$\left\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\right\|_2 = \max\left\{\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2, \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2\right\}. \tag{5.15.8}$$

But $\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2$ because we can again use (5.2.6) to write

$$\begin{aligned}
\frac{1}{\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2^2} &= \min_{\|\mathbf{x}\|_2=1}\left\|\mathbf{U}_1^T\mathbf{V}_1\mathbf{x}\right\|_2^2 = \min_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{V}_1^T\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\mathbf{x} \\
&= \min_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{V}_1^T(\mathbf{I}-\mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\mathbf{x} \\
&= \min_{\|\mathbf{x}\|_2=1}(1-\mathbf{x}^T\mathbf{V}_1^T\mathbf{U}_2\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}) \\
&= 1-\max_{\|\mathbf{x}\|_2=1}\left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2 = 1-\left\|\mathbf{U}_2^T\mathbf{V}_1\right\|_2^2.
\end{aligned}$$

By a similar argument, $1/\left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2^2 = 1-\left\|\mathbf{U}_2^T\mathbf{V}_1\right\|_2^2$ (Exercise 5.15.11(a)). Therefore,

$$\left\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\right\|_2 = \left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|\mathbf{C}^T\right\|_2 = \|\mathbf{C}\|_2 = \|\mathbf{P}_{\mathcal{M}\mathcal{N}}\|_2. \quad \blacksquare$$

While the minimal angle works fine for complementary spaces, it may not convey much information about the separation between noncomplementary subspaces. For example, $\theta_{min} = 0$ whenever $\mathcal{M}$ and $\mathcal{N}$ have a nontrivial intersection, but there nevertheless might be a nontrivial "gap" between $\mathcal{M}$ and $\mathcal{N}$—look at Figure 5.15.1. Rather than thinking about angles to measure such a gap, consider orthogonal distances as discussed in (5.13.13). Define

$$\delta(\mathcal{M},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\left(\mathbf{m},\mathcal{N}\right) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2$$

to be the ***directed distance*** from $\mathcal{M}$ to $\mathcal{N}$, and notice that $\delta(\mathcal{M},\mathcal{N}) \leq 1$ because (5.2.5) and (5.13.10) can be combined to produce

$$dist\,(\mathbf{m},\mathcal{N}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2 = \|\mathbf{P}_{\mathcal{N}^\perp}\mathbf{m}\|_2 \leq \|\mathbf{P}_{\mathcal{N}^\perp}\|_2\,\|\mathbf{m}\|_2 = 1.$$

Figure 5.15.1 illustrates $\delta(\mathcal{M},\mathcal{N})$ for two planes in $\Re^3$.

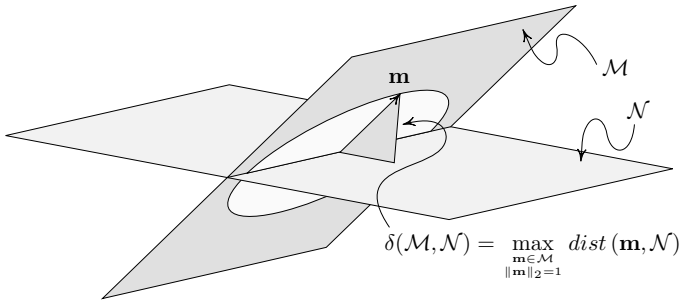


FIGURE 5.15.1

This picture is a bit misleading because $\delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$ for this particular situation. However, $\delta(\mathcal{M},\mathcal{N})$ and $\delta(\mathcal{N},\mathcal{M})$ need not always agree—that's why the phrase *directed* distance is used. For example, if $\mathcal{M}$ is the xy-plane in $\Re^3$ and $\mathcal{N} = span\,\{(0,1,1)\}$, then $\delta(\mathcal{N},\mathcal{M}) = 1/\sqrt{2}$ while $\delta(\mathcal{M},\mathcal{N}) = 1$. Consequently, using orthogonal distance to gauge the degree of maximal separation between an arbitrary pair of subspaces requires that both values of δ be taken into account. Hence we make the following definition.

Gap Between Subspaces

The ***gap*** between subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ is defined to be

$$\text{gap}\,(\mathcal{M},\mathcal{N}) = \max\left\{\delta(\mathcal{M},\mathcal{N}),\ \delta(\mathcal{N},\mathcal{M})\right\}, \tag{5.15.9}$$

where $\delta(\mathcal{M},\mathcal{N}) = \max\limits_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N})$.

Evaluating the gap between a given pair of subspaces requires knowing some properties of directed distance. Observe that (5.15.5) together with the fact that $\|\mathbf{A}^T\|_2 = \|\mathbf{A}\|_2$ can be used to write

$$\begin{aligned}\delta(\mathcal{M},\mathcal{N}) &= \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} dist\,(\mathbf{m},\mathcal{N}) = \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2=1}} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2 \\ &= \max_{\substack{\mathbf{m}\in\mathcal{M}\\ \|\mathbf{m}\|_2\leq 1}} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{m}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\mathbf{x}\|_2 \\ &= \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2 = \|\mathbf{P}_{\mathcal{M}}(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\|_2\,.\end{aligned} \tag{5.15.10}$$

Similarly, $\delta(\mathcal{N},\mathcal{M}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = \|\mathbf{P}_{\mathcal{N}}(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\|_2$. If $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ are the orthogonal matrices introduced on p. 451, then

$$\delta(\mathcal{M},\mathcal{N}) = \|\mathbf{P}_{\mathcal{M}}(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\|_2 = \big\|\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\mathbf{V}_1^T\big\|_2 = \big\|\mathbf{U}_1^T\mathbf{V}_1\big\|_2$$

and (5.15.11)

$$\delta(\mathcal{N},\mathcal{M}) = \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = \big\|\mathbf{U}_2\mathbf{U}_2^T\mathbf{V}_2\mathbf{V}_2^T\big\|_2 = \big\|\mathbf{U}_2^T\mathbf{V}_2\big\|_2 .$$

Combining these observations with (5.15.7) leads us to conclude that

$$\begin{aligned}\|\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}\|_2 &= \max\Big\{\big\|\mathbf{U}_1^T\mathbf{V}_1\big\|_2 , \big\|\mathbf{U}_2^T\mathbf{V}_2\big\|_2\Big\} \\ &= \max\big\{\delta(\mathcal{M},\mathcal{N}),\ \delta(\mathcal{N},\mathcal{M})\big\} \\ &= \operatorname{gap}(\mathcal{M},\mathcal{N}).\end{aligned} \tag{5.15.12}$$

Below is a summary of these and other properties of the gap measure.

Gap Properties

The following statements are true for subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$.

- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = \|\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}\|_2$.
- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = \max\big\{\|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2 , \|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2\big\}$.
- $\operatorname{gap}(\mathcal{M},\mathcal{N}) = 1$ whenever $\dim\mathcal{M} \neq \dim\mathcal{N}$. (5.15.13)
- If $\dim\mathcal{M} = \dim\mathcal{N}$, then $\delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$, and
  - ▷ $\operatorname{gap}(\mathcal{M},\mathcal{N}) = 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $\neq \mathbf{0}$, (5.15.14)
  - ▷ $\operatorname{gap}(\mathcal{M},\mathcal{N}) < 1$ when $\mathcal{M}^\perp \cap \mathcal{N}$ (or $\mathcal{M} \cap \mathcal{N}^\perp$) $= \mathbf{0}$. (5.15.15)

Proof of (5.15.13). Suppose that $\dim\mathcal{M} = r$ and $\dim\mathcal{N} = k$, where $r < k$. Notice that this implies that $\mathcal{M}^\perp \cap \mathcal{N} \neq \mathbf{0}$, for otherwise the formula for the dimension of a sum (4.4.19) yields

$$n \geq \dim(\mathcal{M}^\perp + \mathcal{N}) = \dim\mathcal{M}^\perp + \dim\mathcal{N} = n - r + k > n,$$

which is impossible. Thus there exists a nonzero vector $\mathbf{x} \in \mathcal{M}^\perp \cap \mathcal{N}$, and by normalization we can take $\|\mathbf{x}\|_2 = 1$. Consequently, $(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{x} = \mathbf{x} = \mathbf{P}_{\mathcal{N}}\mathbf{x}$, so $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\mathbf{x}\|_2 = 1$. This insures that $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2 = 1$, which implies $\delta(\mathcal{N},\mathcal{M}) = 1$. ∎

Proof of (5.15.14). Assume $\dim\mathcal{M} = \dim\mathcal{N} = r$, and use the formula for the dimension of a sum along with $(\mathcal{M} \cap \mathcal{N}^\perp)^\perp = \mathcal{M}^\perp + \mathcal{N}$ (Exercise 5.11.5) to conclude that

$$\begin{aligned}\dim\big(\mathcal{M}^\perp \cap \mathcal{N}\big) &= \dim\mathcal{M}^\perp + \dim\mathcal{N} - \dim\big(\mathcal{M}^\perp + \mathcal{N}\big) \\ &= (n-r) + r - \dim\big(\mathcal{M} \cap \mathcal{N}^\perp\big)^\perp = \dim\big(\mathcal{M} \cap \mathcal{N}^\perp\big).\end{aligned}$$

When $\dim\left(\mathcal{M}\cap\mathcal{N}^{\perp}\right)=\dim\left(\mathcal{M}^{\perp}\cap\mathcal{N}\right)>0$, there are vectors $\mathbf{x}\in\mathcal{M}^{\perp}\cap\mathcal{N}$ and $\mathbf{y}\in\mathcal{M}\cap\mathcal{N}^{\perp}$ such that $\|\mathbf{x}\|_2=1=\|\mathbf{y}\|_2$. Hence, $\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\mathbf{x}\|_2=\|\mathbf{x}\|_2=1$, and $\|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\mathbf{y}\|_2=\|\mathbf{y}\|_2=1$, so

$$\delta(\mathcal{N},\mathcal{M})=\|(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\|_2=1=\|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2=\delta(\mathcal{M},\mathcal{N}). \quad \blacksquare$$

Proof of (5.15.15). If $\dim\left(\mathcal{M}\cap\mathcal{N}^{\perp}\right)=\dim\left(\mathcal{M}^{\perp}\cap\mathcal{N}\right)=0$, then $\mathbf{U}_2^T\mathbf{V}_1$ is nonsingular because it is $r\times r$ and has rank r—apply the formula (4.5.1) for the rank of a product. From (5.15.11) we have

$$\begin{aligned}\delta^2(\mathcal{M},\mathcal{N})&=\left\|\mathbf{U}_1^T\mathbf{V}_1\right\|_2^2=\left\|\mathbf{U}_1\mathbf{U}_1^T\mathbf{V}_1\right\|_2^2=\left\|(\mathbf{I}-\mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\right\|_2^2\\&=\max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{V}_1^T(\mathbf{I}-\mathbf{U}_2\mathbf{U}_2^T)\mathbf{V}_1\mathbf{x}=\max_{\|\mathbf{x}\|_2=1}\left(1-\left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2\right)\\&=1-\min_{\|\mathbf{x}\|_2=1}\left\|\mathbf{U}_2^T\mathbf{V}_1\mathbf{x}\right\|_2^2=1-\frac{1}{\left\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\right\|_2^2}<1\ \text{(recall (5.2.6))}.\end{aligned}$$

A similar argument shows $\delta^2(\mathcal{N},\mathcal{M})=\left\|\mathbf{U}_2^T\mathbf{V}_2\right\|_2^2=1-1/\left\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\right\|_2^2$ (Exercise 5.15.11(b)), so $\delta(\mathcal{N},\mathcal{M})=\delta(\mathcal{M},\mathcal{N})<1$. $\blacksquare$

Because $0\le\operatorname{gap}(\mathcal{M},\mathcal{N})\le 1$, the gap measure defines another angle between $\mathcal{M}$ and $\mathcal{N}$.

Maximal Angle

The ***maximal angle*** between subspaces $\mathcal{M},\mathcal{N}\subseteq\Re^n$ is defined to be the number $0\le\theta_{max}\le\pi/2$ for which

$$\sin\theta_{max}=\operatorname{gap}(\mathcal{M},\mathcal{N})=\|\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{N}}\|_2. \tag{5.15.16}$$

For applications requiring knowledge of the degree of separation between a pair of nontrivial complementary subspaces, the minimal angle does the job. Similarly, the maximal angle adequately handles the task for subspaces of equal dimension. However, neither the minimal nor maximal angle may be of much help for more general subspaces. For example, if $\mathcal{M}$ and $\mathcal{N}$ are subspaces of unequal dimension that have a nontrivial intersection, then $\theta_{min}=0$ and $\theta_{max}=\pi/2$, but neither of these numbers might convey the desired information. Consequently, it seems natural to try to formulate definitions of "intermediate" angles between θ_{min} and θ_{max}. There are a host of such angles known as the ***principal*** or ***canonical angles***, and they are derived as follows.

Let $k = \min\{\dim \mathcal{M}, \dim \mathcal{N}\}$, and set $\mathcal{M}_1 = \mathcal{M}$, $\mathcal{N}_1 = \mathcal{N}$, and $\theta_1 = \theta_{min}$. Let $\mathbf{u}_1$ and $\mathbf{v}_1$ be vectors of unit 2-norm such that the following maximum is attained when $\mathbf{u} = \mathbf{u}_1$ and $\mathbf{v} = \mathbf{v}_1$:

$$\cos\theta_{min} = \max_{\substack{\mathbf{u}\in\mathcal{M},\, \mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_1^T\mathbf{u}_1.$$

Set

$$\mathcal{M}_2 = \mathbf{u}_1^\perp \cap \mathcal{M}_1 \quad \text{and} \quad \mathcal{N}_2 = \mathbf{v}_1^\perp \cap \mathcal{N}_1,$$

and define the second principal angle θ_2 to be the minimal angle between $\mathcal{M}_2$ and $\mathcal{N}_2$. Continue in this manner—e.g., if $\mathbf{u}_2$ and $\mathbf{v}_2$ are vectors such that $\|\mathbf{u}_2\|_2 = 1 = \|\mathbf{v}_2\|_2$ and

$$\cos\theta_2 = \max_{\substack{\mathbf{u}\in\mathcal{M}_2,\, \mathbf{v}\in\mathcal{N}_2 \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_2^T\mathbf{u}_2,$$

set

$$\mathcal{M}_3 = \mathbf{u}_2^\perp \cap \mathcal{M}_2 \quad \text{and} \quad \mathcal{N}_3 = \mathbf{v}_2^\perp \cap \mathcal{N}_2,$$

and define the third principal angle θ_3 to be the minimal angle between $\mathcal{M}_3$ and $\mathcal{N}_3$. This process is repeated k times, at which point one of the subspaces is zero. Below is a summary.

Principal Angles

For nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$ with $k = \min\{\dim \mathcal{M}, \dim \mathcal{N}\}$, the principal angles between $\mathcal{M} = \mathcal{M}_1$ and $\mathcal{N} = \mathcal{N}_1$ are recursively defined to be the numbers $0 \le \theta_i \le \pi/2$ such that

$$\cos\theta_i = \max_{\substack{\mathbf{u}\in\mathcal{M}_i,\, \mathbf{v}\in\mathcal{N}_i \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = \mathbf{v}_i^T\mathbf{u}_i, \quad i = 1, 2, \ldots, k,$$

where $\|\mathbf{u}_i\|_2 = 1 = \|\mathbf{v}_i\|_2$, $\mathcal{M}_i = \mathbf{u}_{i-1}^\perp \cap \mathcal{M}_{i-1}$, and $\mathcal{N}_i = \mathbf{v}_{i-1}^\perp \cap \mathcal{N}_{i-1}$.

- It's possible to prove that $\theta_{min} = \theta_1 \le \theta_2 \le \cdots \le \theta_k \le \theta_{max}$, where $\theta_k = \theta_{max}$ when $\dim \mathcal{M} = \dim \mathcal{N}$.
- The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are not uniquely defined, but the θ_i's are unique. In fact, it can be proven that the $\sin\theta_i$'s are singular values (p. 412) for $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$. Furthermore, if $\dim \mathcal{M} \ge \dim \mathcal{N} = k$, then the $\cos\theta_i$'s are the singular values of $\mathbf{V}_2^T\mathbf{U}_1$, and the $\sin\theta_i$'s are the singular values of $\mathbf{V}_2^T\mathbf{U}_2\mathbf{U}_2^T$, where $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ are the orthogonal matrices from p. 451.

Exercises for section 5.15

5.15.1. Determine the angles θ_{min} and θ_{max} between the following subspaces of $\Re^3$.

(a) $\mathcal{M} =$ xy-plane, $\mathcal{N} = span\{(1, 0, 0), (0, 1, 1)\}$.

(b) $\mathcal{M} =$ xy-plane, $\mathcal{N} = span\{(0, 1, 1)\}$.

5.15.2. Determine the principal angles between the following subspaces of $\Re^3$.

(a) $\mathcal{M} =$ xy-plane, $\mathcal{N} = span\{(1, 0, 0), (0, 1, 1)\}$.

(b) $\mathcal{M} =$ xy-plane, $\mathcal{N} = span\{(0, 1, 1)\}$.

5.15.3. Let θ_{min} be the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$.

(a) Explain why $\theta_{max} = 0$ if and only if $\mathcal{M} = \mathcal{N}$.

(b) Explain why $\theta_{min} = 0$ if and only if $\mathcal{M} \cap \mathcal{N} \neq \mathbf{0}$.

(c) Explain why $\theta_{min} = \pi/2$ if and only if $\mathcal{M} \perp \mathcal{N}$.

5.15.4. Let θ_{min} be the minimal angle between nonzero subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, and let $\theta^{\perp}_{min}$ denote the minimal angle between $\mathcal{M}^{\perp}$ and $\mathcal{N}^{\perp}$. Prove that if $\mathcal{M} \oplus \mathcal{N} = \Re^n$, then $\theta_{min} = \theta^{\perp}_{min}$.

5.15.5. For nonzero subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let $\tilde{\theta}_{min}$ denote the minimal angle between $\mathcal{M}$ and $\mathcal{N}^{\perp}$, and let θ_{max} be the maximal angle between $\mathcal{M}$ and $\mathcal{N}$. Prove that if $\mathcal{M} \oplus \mathcal{N}^{\perp} = \Re^n$, then $\cos\tilde{\theta}_{min} = \sin\theta_{max}$.

5.15.6. For subspaces $\mathcal{M}, \mathcal{N} \subseteq \Re^n$, prove that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular if and only if $\mathcal{M}$ and $\mathcal{N}$ are complementary.

5.15.7. For complementary spaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let $\mathbf{P} = \mathbf{P}_{\mathcal{M}\mathcal{N}}$ be the oblique projector onto $\mathcal{M}$ along $\mathcal{N}$, and let $\mathbf{Q} = \mathbf{P}_{\mathcal{M}^{\perp}\mathcal{N}^{\perp}}$ be the oblique projector onto $\mathcal{M}^{\perp}$ along $\mathcal{N}^{\perp}$.

(a) Prove that $(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1} = \mathbf{P} - \mathbf{Q}$.

(b) If θ_{min} is the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, explain why

$$\sin\theta_{min} = \frac{1}{\|\mathbf{P} - \mathbf{Q}\|_2}.$$

(c) Explain why $\|\mathbf{P} - \mathbf{Q}\|_2 = \|\mathbf{P}\|_2$.

5.15.8. Prove that if $f : \mathcal{V} \to \Re$ is a function defined on a space $\mathcal{V}$ such that $f(\alpha\mathbf{x}) = \alpha f(\mathbf{x})$ for scalars $\alpha \geq 0$, then

$$\max_{\|\mathbf{x}\|=1} f(\mathbf{x}) = \max_{\|\mathbf{x}\|\leq 1} f(\mathbf{x}).$$

5.15.9. Let $\mathcal{M}$ and $\mathcal{N}$ be nonzero complementary subspaces of $\Re^n$.

(a) Explain why $\mathbf{P}_{\mathcal{MN}} = \big[(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\big]^{\dagger}$, where $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, and $\mathbf{P}_{\mathcal{MN}}$ is the *oblique* projector onto $\mathcal{M}$ along $\mathcal{N}$.

(b) If θ_{min} is the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, explain why

$$\begin{aligned}\sin\theta_{min} &= \Big\|\big[(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\big]^{\dagger}\Big\|_2^{-1} = \Big\|\big[\mathbf{P}_{\mathcal{M}}(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\big]^{\dagger}\Big\|_2^{-1} \\ &= \Big\|\big[(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\big]^{\dagger}\Big\|_2^{-1} = \Big\|\big[\mathbf{P}_{\mathcal{N}}(\mathbf{I} - \mathbf{P}_{\mathcal{M}})\big]^{\dagger}\Big\|_2^{-1}.\end{aligned}$$

5.15.10. For complementary subspaces $\mathcal{M}, \mathcal{N} \subset \Re^n$, let θ_{min} be the minimal angle between $\mathcal{M}$ and $\mathcal{N}$, and let $\bar{\theta}_{min}$ denote the minimal angle between $\mathcal{M}$ and $\mathcal{N}^{\perp}$.

(a) If $\mathbf{P}_{\mathcal{MN}}$ is the oblique projector onto $\mathcal{M}$ along $\mathcal{N}$, prove that

$$\cos\bar{\theta}_{min} = \Big\|\mathbf{P}_{\mathcal{MN}}^{\dagger}\Big\|_2.$$

(b) Explain why $\sin\theta_{min} \leq \cos\bar{\theta}_{min}$.

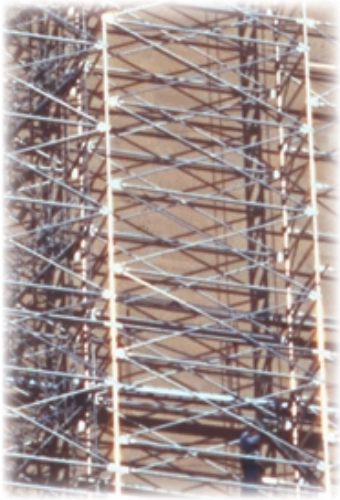
5.15.11. Let $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$ be the orthogonal matrices defined on p. 451.

(a) Prove that if $\mathbf{U}_2^T\mathbf{V}_2$ is nonsingular, then

$$\frac{1}{\big\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\big\|_2^2} = 1 - \big\|\mathbf{U}_2^T\mathbf{V}_1\big\|_2^2.$$

(b) Prove that if $\mathbf{U}_2^T\mathbf{V}_1$ is nonsingular, then

$$\big\|\mathbf{U}_2^T\mathbf{V}_2\big\|_2^2 = 1 - \frac{1}{\big\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\big\|_2^2}.$$



CHAPTER 6

Determinants

6.1 DETERMINANTS

At the beginning of this text, reference was made to the ancient Chinese counting board on which colored bamboo rods were manipulated according to prescribed "rules of thumb" in order to solve a system of linear equations. The Chinese counting board is believed to date back to at least 200 B.C., and it was used more or less in the same way for a millennium. The counting board and the "rules of thumb" eventually found their way to Japan where Seki Kowa (1642–1708), a great Japanese mathematician, synthesized the ancient Chinese ideas of array manipulation. Kowa formulated the concept of what we now call the determinant to facilitate solving linear systems—his definition is thought to have been made some time before 1683.

About the same time—somewhere between 1678 and 1693—Gottfried W. Leibniz (1646–1716), a German mathematician, was independently developing his own concept of the determinant together with applications of array manipulation to solve systems of linear equations. It appears that Leibniz's early work dealt with only three equations in three unknowns, whereas Seki Kowa gave a general treatment for n equations in n unknowns. It seems that Kowa and Leibniz both developed what later became known as Cramer's rule (p. 476), but not in the same form or notation. These men had something else in common—their ideas concerning the solution of linear systems were never adopted by the mathematical community of their time, and their discoveries quickly faded into oblivion.

Eventually the determinant was rediscovered, and much was written on the subject between 1750 and 1900. During this era, determinants became the major tool used to analyze and solve linear systems, while the theory of matrices remained relatively undeveloped. But mathematics, like a river, is everchanging

in its course, and major branches can dry up to become minor tributaries while small trickling brooks can develop into raging torrents. This is precisely what occurred with determinants and matrices. The study and use of determinants eventually gave way to Cayley's matrix algebra, and today matrix and linear algebra are in the main stream of applied mathematics, while the role of determinants has been relegated to a minor backwater position. Nevertheless, it is still important to understand what a determinant is and to learn a few of its fundamental properties. Our goal is not to study determinants for their own sake, but rather to explore those properties that are useful in the further development of matrix theory and its applications. Accordingly, many secondary properties are omitted or confined to the exercises, and the details in proofs will be kept to a minimum.

Over the years there have evolved various "slick" ways to define the determinant, but each of these "slick" approaches seems to require at least one "sticky" theorem in order to make the theory sound. We are going to opt for expedience over elegance and proceed with the classical treatment.

A ***permutation*** $p = (p_1, p_2, \ldots, p_n)$ of the numbers $(1, 2, \ldots, n)$ is simply any rearrangement. For example, the set

$$\{(1,2,3) \quad (1,3,2) \quad (2,1,3) \quad (2,3,1) \quad (3,1,2) \quad (3,2,1)\}$$

contains the six distinct permutations of $(1, 2, 3)$. In general, the sequence $(1, 2, \ldots, n)$ has $n! = n(n-1)(n-2)\cdots 1$ different permutations. Given a permutation, consider the problem of restoring it to natural order by a sequence of pairwise interchanges. For example, $(1, 4, 3, 2)$ can be restored to natural order with a single interchange of 2 and 4 or, as indicated in Figure 6.1.1, three *adjacent* interchanges can be used.

FIGURE 6.1.1

The important thing here is that both 1 and 3 are odd. Try to restore $(1, 4, 3, 2)$ to natural order by using an even number of interchanges, and you will discover that it is impossible. This is due to the following general rule that is stated without proof. *The* ***parity*** *of a permutation is unique*—i.e., if a permutation p can be restored to natural order by an even (odd) number of interchanges, then every other sequence of interchanges that restores p to natural order must

also be even (odd). Accordingly, the ***sign of a permutation*** p is defined to be the number

$$\sigma(p) = \begin{cases} +1 & \text{if } p \text{ can be restored to natural order by an} \\ & \textit{even} \text{ number of interchanges,} \\ -1 & \text{if } p \text{ can be restored to natural order by an} \\ & \textit{odd} \text{ number of interchanges.} \end{cases}$$

For example, if $p = (1, 4, 3, 2)$, then $\sigma(p) = -1$, and if $p = (4, 3, 2, 1)$, then $\sigma(p) = +1$. The sign of the natural order $p = (1, 2, 3, 4)$ is naturally $\sigma(p) = +1$. The general definition of the determinant can now be given.

Definition of Determinant

For an $n \times n$ matrix $\mathbf{A} = [a_{ij}]$, the ***determinant*** of $\mathbf{A}$ is defined to be the scalar

$$\det(\mathbf{A}) = \sum_p \sigma(p) a_{1p_1} a_{2p_2} \cdots a_{np_n}, \tag{6.1.1}$$

where the sum is taken over the $n!$ permutations $p = (p_1, p_2, \ldots, p_n)$ of $(1, 2, \ldots, n)$. Observe that each term $a_{1p_1} a_{2p_2} \cdots a_{np_n}$ in (6.1.1) contains exactly one entry from each row and each column of $\mathbf{A}$. The determinant of $\mathbf{A}$ can be denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$, whichever is more convenient.

Note: The determinant of a nonsquare matrix is not defined.

For example, when $\mathbf{A}$ is 2×2 there are $2! = 2$ permutations of (1,2), namely, $\{(1, 2) \quad (2, 1)\}$, so $\det(\mathbf{A})$ contains the two terms

$$\sigma(1, 2) a_{11} a_{22} \quad \text{and} \quad \sigma(2, 1) a_{12} a_{21}.$$

Since $\sigma(1, 2) = +1$ and $\sigma(2, 1) = -1$, we obtain the familiar formula

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11} a_{22} - a_{12} a_{21}. \tag{6.1.2}$$

Example 6.1.1

Problem: Use the definition to compute $\det(\mathbf{A})$, where $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.

Solution: The $3! = 6$ permutations of $(1, 2, 3)$ together with the terms in the expansion of $\det(\mathbf{A})$ are shown in Table 6.1.1.

Table 6.1.1

$p = (p_1, p_2, p_3)$	$\sigma(p)$	$a_{1p_1}a_{2p_2}a_{3p_3}$
$(1, 2, 3)$	$+$	$1 \times 5 \times 9 = 45$
$(1, 3, 2)$	$-$	$1 \times 6 \times 8 = 48$
$(2, 1, 3)$	$-$	$2 \times 4 \times 9 = 72$
$(2, 3, 1)$	$+$	$2 \times 6 \times 7 = 84$
$(3, 1, 2)$	$+$	$3 \times 4 \times 8 = 96$
$(3, 2, 1)$	$-$	$3 \times 5 \times 7 = 105$

Therefore,

$$\det(\mathbf{A}) = \sum_p \sigma(p) a_{1p_1} a_{2p_2} a_{3p_3} = 45 - 48 - 72 + 84 + 96 - 105 = 0.$$

Perhaps you have seen rules for computing 3×3 determinants that involve running up, down, and around various diagonal lines. These rules do not easily generalize to matrices of order greater than three, and in case you have forgotten (or never knew) them, do not worry about it. Remember the 2×2 rule given in (6.1.2) as well as the following statement concerning triangular matrices and let it go at that.

Triangular Determinants

The determinant of a triangular matrix is the product of its diagonal entries. In other words,

$$\begin{vmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{vmatrix} = t_{11}t_{22}\cdots t_{nn}. \tag{6.1.3}$$

Proof. Recall from the definition (6.1.1) that each term $t_{1p_1}t_{2p_2}\cdots t_{np_n}$ contains exactly one entry from each row and each column. This means that there is only one term in the expansion of the determinant that does not contain an entry below the diagonal, and this term is $t_{11}t_{22}\cdots t_{nn}$. ∎

Transposition Doesn't Alter Determinants

- $\det\left(\mathbf{A}^T\right) = \det(\mathbf{A})$ for all $n \times n$ matrices. (6.1.4)

Proof. As $p = (p_1, p_2, \ldots, p_n)$ varies over all permutations of $(1, 2, \ldots, n)$, the set of all products $\{\sigma(p)a_{1p_1}a_{2p_2}\cdots a_{np_n}\}$ is the same as the set of all products $\{\sigma(p)a_{p_1 1}a_{p_2 2}\cdots a_{p_n n}\}$. Explicitly construct both of these sets for $n = 3$ to convince yourself. ∎

Equation (6.1.4) insures that it's not necessary to distinguish between rows and columns when discussing properties of determinants, so theorems concerning determinants that involve row manipulations will remain true when the word "row" is replaced by "column." For example, it's essential to know how elementary row and column operations alter the determinant of a matrix, but, by virtue of (6.1.4), it suffices to limit the discussion to elementary row operations.

Effects of Row Operations

Let $\mathbf{B}$ be the matrix obtained from $\mathbf{A}_{n\times n}$ by one of the three elementary row operations:

Type I: Interchange rows i and j.

Type II: Multiply row i by $\alpha \neq 0$.

Type III: Add α times row i to row j.

The value of $\det(\mathbf{B})$ is as follows:

- $\det(\mathbf{B}) = -\det(\mathbf{A})$ for Type I operations. (6.1.5)
- $\det(\mathbf{B}) = \alpha\det(\mathbf{A})$ for Type II operations. (6.1.6)
- $\det(\mathbf{B}) = \det(\mathbf{A})$ for Type III operations. (6.1.7)

Proof of (6.1.5). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{i*} = \mathbf{A}_{j*}$ and $\mathbf{B}_{j*} = \mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$ of $(1, 2, \ldots, n)$,

$$\begin{aligned} b_{1p_1}\cdots b_{ip_i}\cdots b_{jp_j}\cdots b_{np_n} &= a_{1p_1}\cdots a_{jp_i}\cdots a_{ip_j}\cdots a_{np_n} \\ &= a_{1p_1}\cdots a_{ip_j}\cdots a_{jp_i}\cdots a_{np_n}. \end{aligned}$$

Furthermore, $\sigma(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = -\sigma(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n)$ because the two permutations differ only by one interchange. Consequently, definition (6.1.1) of the determinant guarantees that $\det(\mathbf{B}) = -\det(\mathbf{A})$.

Proof of (6.1.6). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{i*} = \alpha\mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$,

$$b_{1p_1} \cdots b_{ip_i} \cdots b_{np_n} = a_{1p_1} \cdots \alpha a_{ip_i} \cdots a_{np_n} = \alpha(a_{1p_1} \cdots a_{ip_i} \cdots a_{np_n}),$$

and therefore the expansion (6.1.1) yields $\det(\mathbf{B}) = \alpha \det(\mathbf{A})$.

Proof of (6.1.7). If $\mathbf{B}$ agrees with $\mathbf{A}$ except that $\mathbf{B}_{j*} = \mathbf{A}_{j*} + \alpha\mathbf{A}_{i*}$, then for each permutation $p = (p_1, p_2, \ldots, p_n)$,

$$\begin{aligned} b_{1p_1} \cdots b_{ip_i} \cdots b_{jp_j} \cdots b_{np_n} &= a_{1p_1} \cdots a_{ip_i} \cdots (a_{jp_j} + \alpha a_{ip_j}) \cdots a_{np_n} \\ &= a_{1p_1} \cdots a_{ip_i} \cdots a_{jp_j} \cdots a_{np_n} + \alpha(a_{1p_1} \cdots a_{ip_i} \cdots a_{ip_j} \cdots a_{np_n}), \end{aligned}$$

so that

$$\begin{aligned} \det(\mathbf{B}) = &\sum_p \sigma(p) a_{1p_1} \cdots a_{ip_i} \cdots a_{jp_j} \cdots a_{np_n} \\ &+ \alpha \sum_p \sigma(p) a_{1p_1} \cdots a_{ip_i} \cdots a_{ip_j} \cdots a_{np_n}. \end{aligned} \tag{6.1.8}$$

The first sum on the right-hand side of (6.1.8) is $\det(\mathbf{A})$, while the second sum is the expansion of the determinant of a matrix $\tilde{\mathbf{A}}$ in which the ith and j^{th} rows are identical. For such a matrix, $\det(\tilde{\mathbf{A}}) = 0$ because (6.1.5) says that the sign of the determinant is reversed whenever the ith and j^{th} rows are interchanged, so $\det(\tilde{\mathbf{A}}) = -\det(\tilde{\mathbf{A}})$. Consequently, the second sum on the right-hand side of (6.1.8) is zero, and thus $\det(\mathbf{B}) = \det(\mathbf{A})$. ∎

It is now possible to evaluate the determinant of an elementary matrix associated with any of the three types of elementary operations. Let $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ be elementary matrices of Types I, II, and III, respectively, and recall from the discussion in §3.9 that each of these elementary matrices can be obtained by performing the associated row (or column) operation to an identity matrix of appropriate size. The result concerning triangular determinants (6.1.3) guarantees that $\det(\mathbf{I}) = 1$ regardless of the size of $\mathbf{I}$, so if $\mathbf{E}$ is obtained by interchanging any two rows (or columns) in $\mathbf{I}$, then (6.1.5) insures that

$$\det(\mathbf{E}) = -\det(\mathbf{I}) = -1. \tag{6.1.9}$$

Similarly, if $\mathbf{F}$ is obtained by multiplying any row (or column) in $\mathbf{I}$ by $\alpha \neq 0$, then (6.1.6) implies that

$$\det(\mathbf{F}) = \alpha \det(\mathbf{I}) = \alpha, \tag{6.1.10}$$

and if $\mathbf{G}$ is the result of adding a multiple of one row (or column) in $\mathbf{I}$ to another row (or column) in $\mathbf{I}$, then (6.1.7) guarantees that

$$\det(\mathbf{G}) = \det(\mathbf{I}) = 1. \tag{6.1.11}$$

In particular, (6.1.9)–(6.1.11) guarantee that the determinants of elementary matrices of Types I, II, and III are nonzero.

As discussed in §3.9, if $\mathbf{P}$ is an elementary matrix of Type I, II, or III, and if $\mathbf{A}$ is any other matrix, then the product $\mathbf{PA}$ is the matrix obtained by performing the elementary operation associated with $\mathbf{P}$ to the rows of $\mathbf{A}$. This, together with the observations (6.1.5)–(6.1.7) and (6.1.9)–(6.1.11), leads to the conclusion that for every square matrix $\mathbf{A}$,

$$\begin{aligned}
\det(\mathbf{EA}) &= -\det(\mathbf{A}) &&= \det(\mathbf{E})\det(\mathbf{A}),\\
\det(\mathbf{FA}) &= \alpha\det(\mathbf{A}) &&= \det(\mathbf{F})\det(\mathbf{A}),\\
\det(\mathbf{GA}) &= \det(\mathbf{A}) &&= \det(\mathbf{G})\det(\mathbf{A}).
\end{aligned}$$

In other words, $\det(\mathbf{PA}) = \det(\mathbf{P})\det(\mathbf{A})$ whenever $\mathbf{P}$ is an elementary matrix of Type I, II, or III. It's easy to extend this observation to any number of these elementary matrices, $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$, by writing

$$\begin{aligned}
\det(\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A}) &= \det(\mathbf{P}_1)\det(\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A})\\
&= \det(\mathbf{P}_1)\det(\mathbf{P}_2)\det(\mathbf{P}_3\cdots\mathbf{P}_k\mathbf{A})\\
&\;\;\vdots\\
&= \det(\mathbf{P}_1)\det(\mathbf{P}_2)\cdots\det(\mathbf{P}_k)\det(\mathbf{A}).
\end{aligned} \tag{6.1.12}$$

This leads to a characterization of invertibility in terms of determinants.

Invertibility and Determinants

- $\mathbf{A}_{n\times n}$ is nonsingular if and only if $\det(\mathbf{A}) \neq 0$ (6.1.13)

  or, equivalently,
- $\mathbf{A}_{n\times n}$ is singular if and only if $\det(\mathbf{A}) = 0$. (6.1.14)

Proof. Let $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_k$ be a sequence of elementary matrices of Type I, II, or III such that $\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{A} = \mathbf{E}_{\mathbf{A}}$, and apply (6.1.12) to conclude

$$\det(\mathbf{P}_1)\det(\mathbf{P}_2)\cdots\det(\mathbf{P}_k)\det(\mathbf{A}) = \det(\mathbf{E}_{\mathbf{A}}).$$

Since elementary matrices have nonzero determinants,

$$\begin{aligned}
\det(\mathbf{A}) \neq 0 &\iff \det(\mathbf{E}_{\mathbf{A}}) \neq 0 \iff \text{ there are no zero pivots}\\
&\iff \text{every column in } \mathbf{E}_{\mathbf{A}} \text{ (and in } \mathbf{A}\text{) is basic}\\
&\iff \mathbf{A} \text{ is nonsingular.} \quad \blacksquare
\end{aligned}$$

Example 6.1.2

Caution! Small Determinants $\not\Longleftrightarrow$ Near Singularity. Because of (6.1.13) and (6.1.14), it might be easy to get the idea that $\det(\mathbf{A})$ is somehow a measure of how close $\mathbf{A}$ is to being singular, but this is not necessarily the case. Nearly singular matrices need not have determinants of small magnitude. For example, $\mathbf{A}_n = \begin{pmatrix} n & 0 \\ 0 & 1/n \end{pmatrix}$ is nearly singular when n is large, but $\det(\mathbf{A}_n) = 1$ for all n. Furthermore, small determinants do not necessarily signal nearly singular matrices. For example,

$$\mathbf{A}_n = \begin{pmatrix} .1 & 0 & \cdots & 0 \\ 0 & .1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & .1 \end{pmatrix}_{n\times n}$$

is not close to any singular matrix—see (5.12.10) on p. 417—but $\det(\mathbf{A}_n) = (.1)^n$ is extremely small for large n.

A ***minor determinant*** (or simply a ***minor***) of $\mathbf{A}_{m\times n}$ is defined to be the determinant of any $k \times k$ submatrix of $\mathbf{A}$. For example,

$$\begin{vmatrix} 1 & 2 \\ 4 & 5 \end{vmatrix} = -3 \text{ and } \begin{vmatrix} 2 & 3 \\ 8 & 9 \end{vmatrix} = -6 \text{ are } 2\times 2 \text{ minors of } \mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

An individual entry of $\mathbf{A}$ can be regarded as a 1×1 minor, and $\det(\mathbf{A})$ itself is considered to be a 3×3 minor of $\mathbf{A}$.

We already know that the rank of any matrix $\mathbf{A}$ is the size of the largest nonsingular submatrix in $\mathbf{A}$ (p. 215). But (6.1.13) guarantees that the nonsingular submatrices of $\mathbf{A}$ are simply those submatrices with nonzero determinants, so we have the following characterization of rank.

Rank and Determinants

- $rank(\mathbf{A})$ = the size of the largest nonzero minor of $\mathbf{A}$.

Example 6.1.3

Problem: Use determinants to compute the rank of $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 1 \\ 7 & 8 & 9 & 1 \end{pmatrix}$.

Solution: Clearly, there are 1×1 and 2×2 minors that are nonzero, so $rank(\mathbf{A}) \geq 2$. In order to decide if the rank is three, we must see if there

are any 3×3 nonzero minors. There are exactly four 3×3 minors, and they are

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

Since all 3×3 minors are 0, we conclude that $rank\,(\mathbf{A}) = 2$. You should be able to see from this example that using determinants is generally not a good way to compute the rank of a matrix.

In (6.1.12) we observed that the determinant of a product of elementary matrices is the product of their respective determinants. We are now in a position to extend this observation.

Product Rules

- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$ for all $n \times n$ matrices. (6.1.15)
- $\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} = \det(\mathbf{A})\det(\mathbf{D})$ if $\mathbf{A}$ and $\mathbf{D}$ are square. (6.1.16)

Proof of (6.1.15). If $\mathbf{A}$ is singular, then $\mathbf{AB}$ is also singular because (4.5.2) says that $rank\,(\mathbf{AB}) \le rank\,(\mathbf{A})$. Consequently, (6.1.14) implies that

$$\det(\mathbf{AB}) = 0 = \det(\mathbf{A})\det(\mathbf{B}),$$

so (6.1.15) is trivially true when $\mathbf{A}$ is singular. If $\mathbf{A}$ is nonsingular, then $\mathbf{A}$ can be written as a product of elementary matrices $\mathbf{A} = \mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k$ that are of Type I, II, or III—recall (3.9.3). Therefore, (6.1.12) can be applied to produce

$$\begin{aligned} \det(\mathbf{AB}) &= \det(\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k\mathbf{B}) = \det(\mathbf{P}_1)\det(\mathbf{P}_2)\cdots\det(\mathbf{P}_k)\det(\mathbf{B}) \\ &= \det(\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_k)\det(\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B}). \end{aligned}$$

Proof of (6.1.16). First consider the special case $\mathbf{X} = \begin{pmatrix} \mathbf{A}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, and use the definition to write $\det(\mathbf{X}) = \sum_{\sigma(p)} x_{1j_1}x_{2j_2}\cdots x_{rj_r}x_{r+1,j_{r+1}}\cdots x_{n,j_n}$. But

$$x_{rj_r}x_{r+1,j_{r+1}}\cdots x_{n,j_n} = \begin{cases} 1 & \text{when } p = \begin{pmatrix} 1 & \cdots & r & r+1 & \cdots & n \\ j_1 & \cdots & j_r & r+1 & \cdots & n \end{pmatrix}, \\ 0 & \text{for all other permutations,} \end{cases}$$

so, if p_r denotes permutations of only the first r positive integers, then

$$\det(\mathbf{X}) = \sum_{\sigma(p)} x_{1j_1}x_{2j_2}\cdots x_{rj_r}x_{r+1,j_{r+1}}\cdots x_{n,j_n} = \sum_{\sigma(p_r)} x_{1j_1}x_{2j_2}\cdots x_{rj_r} = \det(\mathbf{A}).$$

Thus $\begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{vmatrix} = \det(\mathbf{A})$. Similarly, $\begin{vmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det(\mathbf{D})$, so, by (6.1.15),

$$\begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det\left\{\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}\right\} = \begin{vmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{vmatrix}\begin{vmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \det(\mathbf{A})\det(\mathbf{D}).$$

If $\mathbf{A} = \mathbf{Q}_A\mathbf{R}_A$ and $\mathbf{D} = \mathbf{Q}_D\mathbf{R}_D$ are the respective QR factorizations (p. 345) of $\mathbf{A}$ and $\mathbf{D}$, then $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_D \end{pmatrix}\begin{pmatrix} \mathbf{R}_A & \mathbf{Q}_A^T\mathbf{B} \\ \mathbf{0} & \mathbf{R}_D \end{pmatrix}$ is also a QR factorization. By (6.1.3), the determinant of a triangular matrix is the product of its diagonal entries, and this together with the previous results yield

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{D} \end{vmatrix} = \begin{vmatrix} \mathbf{Q}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_D \end{vmatrix}\begin{vmatrix} \mathbf{R}_A & \mathbf{Q}_A^T\mathbf{B} \\ \mathbf{0} & \mathbf{R}_D \end{vmatrix} = \det(\mathbf{Q}_A)\det(\mathbf{Q}_D)\det(\mathbf{R}_A)\det(\mathbf{R}_D)$$
$$= \det(\mathbf{Q}_A\mathbf{R}_A)\det(\mathbf{Q}_D\mathbf{R}_D) = \det(\mathbf{A})\det(\mathbf{D}). \quad \blacksquare$$

Example 6.1.4

Volume and Determinants. The definition of a determinant is purely algebraic, but there is a concrete geometrical interpretation. A solid in $\Re^m$ with parallel opposing faces whose adjacent sides are defined by vectors from a linearly independent set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is called an n-dimensional ***parallelepiped***. As depicted in Figure 6.1.2, a two-dimensional parallelepiped is a parallelogram, and a three-dimensional parallelepiped is a skewed rectangular box.

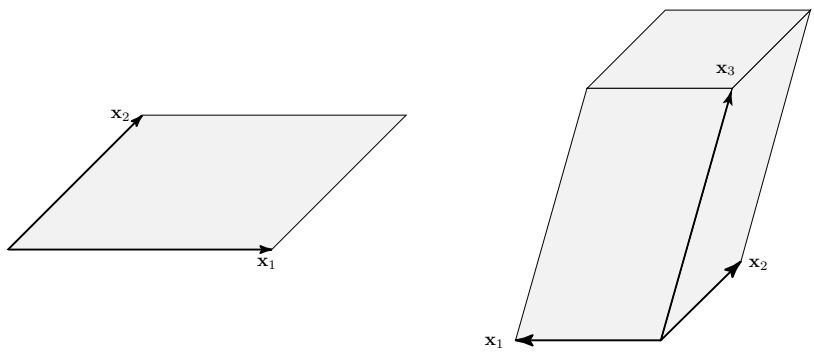


FIGURE 6.1.2

Problem: When $\mathbf{A} \in \Re^{m\times n}$ has linearly independent columns, explain why the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \left[\det\left(\mathbf{A}^T\mathbf{A}\right)\right]^{1/2}$. In particular, if $\mathbf{A}$ is square, then $V_n = |\det(\mathbf{A})|$.

Solution: Recall from Example 5.13.2 on p. 431 that if $\mathbf{A}_{m\times n} = \mathbf{Q}_{m\times n}\mathbf{R}_{n\times n}$ is the (rectangular) QR factorization of $\mathbf{A}$, then the volume of the n-dimensional parallelepiped generated by the columns of $\mathbf{A}$ is $V_n = \nu_1\nu_2\cdots\nu_n = \det(\mathbf{R})$, where the ν_k 's are the diagonal elements of the upper-triangular matrix $\mathbf{R}$. Use

$\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ together with the product rule (6.1.15) and the fact that transposition doesn't affect determinants (6.1.4) to write

$$\begin{aligned}\det\left(\mathbf{A}^T\mathbf{A}\right) &= \det\left(\mathbf{R}^T\mathbf{Q}^T\mathbf{Q}\mathbf{R}\right) = \det\left(\mathbf{R}^T\mathbf{R}\right) = \det\left(\mathbf{R}^T\right)\det\left(\mathbf{R}\right)\\ &= \left(\det\left(\mathbf{R}\right)\right)^2 = \left(\nu_1\nu_2\cdots\nu_n\right)^2 = V_n^2.\end{aligned} \tag{6.1.17}$$

If $\mathbf{A}$ is square, $\det\left(\mathbf{A}^T\mathbf{A}\right) = \det\left(\mathbf{A}^T\right)\det\left(\mathbf{A}\right) = \left(\det\left(\mathbf{A}\right)\right)^2$, so $V_n = |\det\left(\mathbf{A}\right)|$.

Hadamard's Inequality: Recall from (5.13.7) that if

$$\mathbf{A} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\cdots|\, \mathbf{x}_n\right]_{n\times n} \quad \text{and} \quad \mathbf{A}_j = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\cdots|\, \mathbf{x}_j\right]_{n\times j},$$

then $\nu_1 = \|\mathbf{x}_1\|_2$ and $\nu_k = \|(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_k\|_2$ (the projected height of $\mathbf{x}_k$) for $k > 1$, where $\mathbf{P}_k$ is the orthogonal projector onto $R\left(\mathbf{A}_{k-1}\right)$. But

$$\nu_k^2 = \|(\mathbf{I}-\mathbf{P}_k)\mathbf{x}_k\|_2^2 \le \|(\mathbf{I}-\mathbf{P}_k)\|_2^2\,\|\mathbf{x}_k\|_2^2 = \|\mathbf{x}_k\|_2^2 \quad \text{(recall (5.13.10))},$$

so, by (6.1.17), $\det\left(\mathbf{A}^T\mathbf{A}\right) \le \|\mathbf{x}_1\|_2^2\,\|\mathbf{x}_2\|_2^2\cdots\|\mathbf{x}_n\|_2^2$ or, equivalently,

$$|\det\left(\mathbf{A}\right)| \le \prod_{k=1}^{n}\|\mathbf{x}_k\|_2 = \prod_{j=1}^{n}\left(\sum_{i=1}^{n}|a_{ij}|^2\right)^{1/2}, \tag{6.1.18}$$

with equality holding if and only if the $\mathbf{x}_k$'s are mutually orthogonal. This is ***Hadamard's inequality***.[64] In light of the preceding discussion, it simply asserts that the volume of the parallelepiped $\mathcal{P}$ generated by the columns of $\mathbf{A}$ can't exceed the volume of a rectangular box whose sides have length $\|\mathbf{x}_k\|_2$, a fact that is geometrically evident because $\mathcal{P}$ is a *skewed* rectangular box with sides of length $\|\mathbf{x}_k\|_2$.

The product rule (6.1.15) provides a practical way to compute determinants. Recall from §3.10 that for every nonsingular matrix $\mathbf{A}$, there is a permutation matrix $\mathbf{P}$ (which is a product of elementary interchange matrices) such that $\mathbf{PA} = \mathbf{LU}$ in which $\mathbf{L}$ is lower triangular with 1's on its diagonal, and $\mathbf{U}$ is upper triangular with the pivots on its diagonal. The product rule guarantees

[64] Jacques Hadamard (1865–1963), a leading French mathematician of the first half of the twentieth century, discovered this inequality in 1893. Influenced in part by the tragic death of his sons in World War I, Hadamard became a peace activist whose politics drifted far left to the extent that the United States was reluctant to allow him to enter the country to attend the International Congress of Mathematicians held in Cambridge, Massachusetts, in 1950. Due to support from influential mathematicians, Hadamard was made honorary president of the congress, and the resulting visibility together with pressure from important U.S. scientists forced officials to allow him to attend.

that $\det(\mathbf{P})\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U})$, and we know from (6.1.9) that if $\mathbf{E}$ is an elementary interchange matrix, then $\det(\mathbf{E}) = -1$, so

$$\det(\mathbf{P}) = \begin{cases} +1 & \text{if } \mathbf{P} \text{ is the product of an } \textit{even} \text{ number of interchanges,} \\ -1 & \text{if } \mathbf{P} \text{ is the product of an } \textit{odd} \text{ number of interchanges.} \end{cases}$$

The result concerning triangular determinants (6.1.3) shows that $\det(\mathbf{L}) = 1$ and $\det(\mathbf{U}) = u_{11}u_{22}\cdots u_{nn}$, where the u_{ii}'s are the pivots, so, putting these observations together yields $\det(\mathbf{A}) = \pm u_{11}u_{22}\cdots u_{nn}$, where the sign depends on the number of row interchanges used. Below is a summary.

Computing a Determinant

If $\mathbf{PA}_{n\times n} = \mathbf{LU}$ is an LU factorization obtained with row interchanges (use partial pivoting for numerical stability), then

$$\det(\mathbf{A}) = \sigma u_{11}u_{22}\cdots u_{nn}.$$

The u_{ii}'s are the pivots, and σ is the sign of the permutation. That is,

$$\sigma = \begin{cases} +1 & \text{if an } \textit{even} \text{ number of row interchanges are used,} \\ -1 & \text{if an } \textit{odd} \text{ number of row interchanges are used.} \end{cases}$$

If a zero pivot emerges that cannot be removed (because all entries below the pivot are zero), then $\mathbf{A}$ is singular and $\det(\mathbf{A}) = 0$. Exercise 6.2.18 discusses orthogonal reduction to compute $\det(\mathbf{A})$.

Example 6.1.5

Problem: Use partial pivoting to determine an LU decomposition $\mathbf{PA} = \mathbf{LU}$, and then evaluate the determinant of $\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}$.

Solution: The LU factors of $\mathbf{A}$ were computed in Example 3.10.4 as follows.

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The only modification needed is to keep track of how many row interchanges are used. Reviewing Example 3.10.4 reveals that the pivoting process required three interchanges, so $\sigma = -1$, and hence $\det(\mathbf{A}) = (-1)(4)(5)(-6)(1) = 120$.

It's sometimes necessary to compute the derivative of a determinant whose entries are differentiable functions. The following formula shows how this is done.

Derivative of a Determinant

If the entries in $\mathbf{A}_{n\times n} = [a_{ij}(t)]$ are differentiable functions of t, then

$$\frac{d\left(\det(\mathbf{A})\right)}{dt} = \det(\mathbf{D}_1) + \det(\mathbf{D}_2) + \cdots + \det(\mathbf{D}_n), \tag{6.1.19}$$

where $\mathbf{D}_i$ is identical to $\mathbf{A}$ except that the entries in the ith row are replaced by their derivatives—i.e., $[\mathbf{D}_i]_{k*} = \begin{cases} \mathbf{A}_{k*} & \text{if } i \neq k, \\ d\,\mathbf{A}_{k*}/dt & \text{if } i = k. \end{cases}$

Proof. This follows directly from the definition of a determinant by writing

$$\begin{aligned}
\frac{d\big(\det(\mathbf{A})\big)}{dt} &= \frac{d}{dt}\sum_p \sigma(p) a_{1p_1}a_{2p_2}\cdots a_{np_n} = \sum_p \sigma(p)\frac{d\big(a_{1p_1}a_{2p_2}\cdots a_{np_n}\big)}{dt} \\
&= \sum_p \sigma(p)\Big(a'_{1p_1}a_{2p_2}\cdots a_{np_n} + a_{1p_1}a'_{2p_2}\cdots a_{np_n} + \cdots + a_{1p_1}a_{2p_2}\cdots a'_{np_n}\Big) \\
&= \sum_p \sigma(p) a'_{1p_1}a_{2p_2}\cdots a_{np_n} + \sum_p \sigma(p) a_{1p_1}a'_{2p_2}\cdots a_{np_n} \\
&\qquad\qquad + \cdots + \sum_p \sigma(p) a_{1p_1}a_{2p_2}\cdots a'_{np_n} \\
&= \det(\mathbf{D}_1) + \det(\mathbf{D}_2) + \cdots + \det(\mathbf{D}_n). \quad \blacksquare
\end{aligned}$$

Example 6.1.6

Problem: Evaluate the derivative $d\left(\det(\mathbf{A})\right)/dt$ for $\mathbf{A} = \begin{pmatrix} e^t & e^{-t} \\ \cos t & \sin t \end{pmatrix}$.

Solution: Applying formula (6.1.19) yields

$$\frac{d\left(\det(\mathbf{A})\right)}{dt} = \begin{vmatrix} e^t & -e^{-t} \\ \cos t & \sin t \end{vmatrix} + \begin{vmatrix} e^t & e^{-t} \\ -\sin t & \cos t \end{vmatrix} = \left(e^t + e^{-t}\right)(\cos t + \sin t).$$

Check this by first expanding $\det(\mathbf{A})$ and then computing the derivative.

Exercises for section 6.1

6.1.1. Use the definition to evaluate $\det(\mathbf{A})$ for each of the following matrices.

(a) $\mathbf{A} = \begin{pmatrix} 3 & -2 & 1 \\ -5 & 4 & 0 \\ 2 & 1 & 6 \end{pmatrix}$. (b) $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ -2 & 2 & 1 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 0 & 0 & \alpha \\ 0 & \beta & 0 \\ \gamma & 0 & 0 \end{pmatrix}$. (d) $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$.

6.1.2. What is the volume of the parallelepiped generated by the three vectors $\mathbf{x}_1 = (3, 0, -4, 0)^T$, $\mathbf{x}_2 = (0, 2, 0, -2)^T$, and $\mathbf{x}_3 = (0, 1, 0, 1)^T$?

6.1.3. Using Gaussian elimination to reduce $\mathbf{A}$ to an upper-triangular matrix, evaluate $\det(\mathbf{A})$ for each of the following matrices.

(a) $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 1 \\ 1 & 4 & 4 \end{pmatrix}$. (b) $\mathbf{A} = \begin{pmatrix} 1 & 3 & 5 \\ -1 & 4 & 2 \\ 3 & -2 & 4 \end{pmatrix}$.

(c) $\mathbf{A} = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}$. (d) $\mathbf{A} = \begin{pmatrix} 0 & 0 & -2 & 3 \\ 1 & 0 & 1 & 2 \\ -1 & 1 & 2 & 1 \\ 0 & 2 & -3 & 0 \end{pmatrix}$.

(e) $\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$. (f) $\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 3 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n \end{pmatrix}$.

6.1.4. Use determinants to compute the rank of $\mathbf{A} = \begin{pmatrix} 1 & 3 & -2 \\ 0 & 1 & 2 \\ -1 & -1 & 6 \\ 2 & 5 & -6 \end{pmatrix}$.

6.1.5. Use determinants to find the values of α for which the following system possesses a unique solution.

$$\begin{pmatrix} 1 & \alpha & 0 \\ 0 & 1 & -1 \\ \alpha & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \\ 7 \end{pmatrix}.$$

6.1.6. If $\mathbf{A}$ is nonsingular, explain why $\det\left(\mathbf{A}^{-1}\right) = 1/\det(\mathbf{A})$.

6.1.7. Explain why determinants are invariant under similarity transformations. That is, show $\det\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right) = \det(\mathbf{A})$ for all nonsingular $\mathbf{P}$.

6.1.8. Explain why $\det(\mathbf{A}^*) = \overline{\det(\mathbf{A})}$.

6.1.9. (a) Explain why $|\det(\mathbf{Q})| = 1$ when $\mathbf{Q}$ is unitary. In particular, $\det(\mathbf{Q}) = \pm 1$ if $\mathbf{Q}$ is an orthogonal matrix.

(b) How are the singular values of $\mathbf{A} \in \mathcal{C}^{n\times n}$ related to $\det(\mathbf{A})$?

6.1.10. Prove that if $\mathbf{A}$ is $m \times n$, then $\det(\mathbf{A}^*\mathbf{A}) \geq 0$, and explain why $\det(\mathbf{A}^*\mathbf{A}) > 0$ if and only if $rank(\mathbf{A}) = n$.

6.1.11. If $\mathbf{A}$ is $n \times n$, explain why $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$ for all scalars α.

6.1.12. If $\mathbf{A}$ is an $n \times n$ skew-symmetric matrix, prove that $\mathbf{A}$ is singular whenever n is odd. **Hint:** Use Exercise 6.1.11.

6.1.13. How can you build random integer matrices with $\det(\mathbf{A}) = 1$?

6.1.14. If the kth row of $\mathbf{A}_{n\times n}$ is written as a sum $\mathbf{A}_{k*} = \mathbf{x}^T + \mathbf{y}^T + \cdots + \mathbf{z}^T$, where $\mathbf{x^T}, \mathbf{y^T}, \ldots, \mathbf{z^T}$ are row vectors, explain why

$$\det(\mathbf{A}) = \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{x}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix} + \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{y}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix} + \cdots + \det\begin{pmatrix} \mathbf{A}_{1*} \\ \vdots \\ \mathbf{z}^T \\ \vdots \\ \mathbf{A}_{n*} \end{pmatrix}.$$

6.1.15. The CBS inequality (p. 272) says that $|\mathbf{x}^*\mathbf{y}| \leq \|\mathbf{x}\|_2^2 \, \|\mathbf{y}\|_2^2$ for vectors $\mathbf{x}, \mathbf{y} \in \mathcal{C}^{n\times 1}$. Use Exercise 6.1.10 to give an alternate proof of the CBS inequality along with an alternate explanation of why equality holds if and only if $\mathbf{y}$ is a scalar multiple of $\mathbf{x}$.

6.1.16. Determinant Formula for Pivots. Let $\mathbf{A}_k$ be the $k \times k$ leading principal submatrix of $\mathbf{A}_{n\times n}$ (p. 148). Prove that if $\mathbf{A}$ has an LU factorization $\mathbf{A} = \mathbf{LU}$, then $\det(\mathbf{A}_k) = u_{11}u_{22}\cdots u_{kk}$, and deduce that the kth pivot is $u_{kk} = \begin{cases} \det(\mathbf{A}_1) = a_{11} & \text{for } k = 1, \\ \det(\mathbf{A}_k)/\det(\mathbf{A}_{k-1}) & \text{for } k = 2, 3, \ldots, n. \end{cases}$

6.1.17. Prove that if $rank(\mathbf{A}_{m\times n}) = n$, then $\mathbf{A}^T\mathbf{A}$ has an LU factorization with positive pivots—i.e., $\mathbf{A}^T\mathbf{A}$ is *positive definite* (pp. 154 and 559).

6.1.18. Let $\mathbf{A}(x) = \begin{pmatrix} 2-x & 3 & 4 \\ 0 & 4-x & -5 \\ 1 & -1 & 3-x \end{pmatrix}$.

(a) First evaluate $\det(\mathbf{A})$, and then compute $d\big(\det(\mathbf{A})\big)/dx$.

(b) Use formula (6.1.19) to evaluate $d\big(\det(\mathbf{A})\big)/dx$.

6.1.19. When the entries of $\mathbf{A} = [a_{ij}(x)]$ are differentiable functions of x, we define $d\,\mathbf{A}/dx = [d\,a_{ij}/dx]$ (the matrix of derivatives). For square matrices, is it always the case that $d\big(\det(\mathbf{A})\big)/dx = \det(d\mathbf{A}/dx)$?

6.1.20. For a set of functions $\mathcal{S} = \{f_1(x), f_2(x), \ldots, f_n(x)\}$ that are $n-1$ times differentiable, the determinant

$$w(x) = \begin{vmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{vmatrix}$$

is called the ***Wronskian*** of $\mathcal{S}$. If $\mathcal{S}$ is a linearly dependent set, explain why $w(x) = 0$ for every value of x. **Hint:** Recall Example 4.3.6 (p. 189).

6.1.21. Consider evaluating an $n \times n$ determinant from the definition (6.1.1).

(a) How many multiplications are required?

(b) Assuming a computer will do 1,000,000 multiplications per second, and neglecting all other operations, what is the largest order determinant that can be evaluated in one hour?

(c) Under the same conditions of part (b), how long will it take to evaluate the determinant of a 100×100 matrix?
Hint: $100! \approx 9.33 \times 10^{157}$.

(d) If all other operations are neglected, how many multiplications per second must a computer perform if the task of evaluating the determinant of a 100×100 matrix is to be completed in 100 years?

6.2 ADDITIONAL PROPERTIES OF DETERMINANTS

The purpose of this section is to present some additional properties of determinants that will be helpful in later developments.

Block Determinants

If $\mathbf{A}$ and $\mathbf{D}$ are square matrices, then

$$\det\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} = \begin{cases}\det(\mathbf{A})\det\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right) & \text{when } \mathbf{A}^{-1} \text{ exists,}\\ \det(\mathbf{D})\det\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right) & \text{when } \mathbf{D}^{-1} \text{ exists.}\end{cases} \tag{6.2.1}$$

The matrices $\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ are called the ***Schur complements*** of $\mathbf{A}$ and $\mathbf{D}$, respectively—see Exercise 3.7.11 on p. 123.

Proof. If $\mathbf{A}^{-1}$ exists, then $\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{C}\mathbf{A}^{-1} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{0} & \mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\end{pmatrix}$, and the product rules (p. 467) produce the first formula in (6.2.1). The second formula follows by using a similar trick. ∎

Since the determinant of a product is equal to the product of the determinants, it's only natural to inquire if a similar result holds for sums. In other words, is $\det(\mathbf{A}+\mathbf{B}) = \det(\mathbf{A})+\det(\mathbf{B})$? *Almost never*! Try a couple of examples to convince yourself. Nevertheless, there are still some statements that can be made regarding the determinant of certain types of sums. In a loose sense, the result of Exercise 6.1.14 was a statement concerning determinants and sums, but the following result is a little more satisfying.

Rank-One Updates

If $\mathbf{A}_{n\times n}$ is nonsingular, and if $\mathbf{c}$ and $\mathbf{d}$ are $n\times 1$ columns, then

- $\det\left(\mathbf{I}+\mathbf{c}\mathbf{d}^T\right) = 1+\mathbf{d}^T\mathbf{c},$ (6.2.2)
- $\det\left(\mathbf{A}+\mathbf{c}\mathbf{d}^T\right) = \det(\mathbf{A})\left(1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}\right).$ (6.2.3)

Exercise 6.2.7 presents a generalized version of these formulas.

Proof. The proof of (6.2.2) follows by applying the product rules (p. 467) to

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{d}^T & 1\end{pmatrix}\begin{pmatrix}\mathbf{I}+\mathbf{c}\mathbf{d}^T & \mathbf{c}\\ \mathbf{0} & 1\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{d}^T & 1\end{pmatrix} = \begin{pmatrix}\mathbf{I} & \mathbf{c}\\ \mathbf{0} & 1+\mathbf{d}^T\mathbf{c}\end{pmatrix}.$$

To prove (6.2.3), write $\mathbf{A}+\mathbf{c}\mathbf{d}^T = \mathbf{A}\left(\mathbf{I}+\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\right)$, and apply the product rule (6.1.15) along with (6.2.2). ∎

Example 6.2.1

Problem: For $\mathbf{A} = \begin{pmatrix} 1+\lambda_1 & 1 & \cdots & 1 \\ 1 & 1+\lambda_2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1+\lambda_n \end{pmatrix}$, $\lambda_i \neq 0$, find $\det(\mathbf{A})$.

Solution: Express $\mathbf{A}$ as a rank-one updated matrix $\mathbf{A} = \mathbf{D} + \mathbf{e}\mathbf{e}^T$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{e}^T = (1 \quad 1 \quad \cdots \quad 1)$. Apply (6.2.3) to produce

$$\det(\mathbf{D} + \mathbf{e}\mathbf{e}^T) = \det(\mathbf{D})\left(1 + \mathbf{e}^T\mathbf{D}^{-1}\mathbf{e}\right) = \left(\prod_{i=1}^{n} \lambda_i\right)\left(1 + \sum_{i=1}^{n} \frac{1}{\lambda_i}\right).$$

The classical result known as Cramer's rule[65] is a corollary of the rank-one update formula (6.2.3).

Cramer's Rule

In a nonsingular system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$, the ith unknown is

$$x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})},$$

where $\mathbf{A}_i = \left[\mathbf{A}_{*1} \,\middle|\, \cdots \,\middle|\, \mathbf{A}_{*i-1} \,\middle|\, \mathbf{b} \,\middle|\, \mathbf{A}_{*i+1} \,\middle|\, \cdots \,\middle|\, \mathbf{A}_{*n}\right]$. That is, $\mathbf{A}_i$ is identical to $\mathbf{A}$ except that column $\mathbf{A}_{*i}$ has been replaced by $\mathbf{b}$.

Proof. Since $\mathbf{A}_i = \mathbf{A} + (\mathbf{b} - \mathbf{A}_{*i})\,\mathbf{e}_i^T$, where $\mathbf{e}_i$ is the ith unit vector, (6.2.3) may be applied to yield

$$\begin{aligned}\det(\mathbf{A}_i) &= \det(\mathbf{A})\Big(1 + \mathbf{e}_i^T\mathbf{A}^{-1}(\mathbf{b} - \mathbf{A}_{*i})\Big) = \det(\mathbf{A})\Big(1 + \mathbf{e}_i^T(\mathbf{x} - \mathbf{e}_i)\Big) \\ &= \det(\mathbf{A})(1 + x_i - 1) = \det(\mathbf{A})\,x_i.\end{aligned}$$

Thus $x_i = \det(\mathbf{A}_i)/\det(\mathbf{A})$ because $\mathbf{A}$ being nonsingular insures $\det(\mathbf{A}) \neq 0$ by (6.1.13). ∎

[65] Gabriel Cramer (1704–1752) was a mathematician from Geneva, Switzerland. As mentioned in §6.1, Cramer's rule was apparently known to others long before Cramer rediscovered and published it in 1750. Nevertheless, Cramer's recognition is not undeserved because his work was responsible for a revived interest in determinants and systems of linear equations. After Cramer's publication, Cramer's rule met with instant success, and it quickly found its way into the textbooks and classrooms of Europe. It is reported that there was a time when students passed or failed the exams in the schools of public service in France according to their understanding of Cramer's rule.

Example 6.2.2

Problem: Determine the value of t for which $x_3(t)$ is minimized in

$$\begin{pmatrix} t & 0 & 1/t \\ 0 & t & t^2 \\ 1 & t^2 & t^3 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} = \begin{pmatrix} 1 \\ 1/t \\ 1/t^2 \end{pmatrix}.$$

Solution: Only one component of the solution is required, so it's wasted effort to solve the entire system. Use Cramer's rule to obtain

$$x_3(t) = \frac{\begin{vmatrix} t & 0 & 1 \\ 0 & t & 1/t \\ 1 & t^2 & 1/t^2 \end{vmatrix}}{\begin{vmatrix} t & 0 & 1/t \\ 0 & t & t^2 \\ 1 & t^2 & t^3 \end{vmatrix}} = \frac{1-t-t^2}{-1} = t^2+t-1, \quad \text{and set} \quad \frac{d\,x_3(t)}{dt} = 0$$

to conclude that $x_3(t)$ is minimized at $t = -1/2$.

Recall that minor determinants of $\mathbf{A}$ are simply determinants of submatrices of $\mathbf{A}$. We are now in a position to see that in an $n \times n$ matrix the $n-1 \times n-1$ minor determinants have a special significance.

Cofactors

The ***cofactor*** of $\mathbf{A}_{n\times n}$ associated with the (i,j)-position is defined as

$$\mathring{A}_{ij} = (-1)^{i+j} M_{ij},$$

where M_{ij} is the $n-1 \times n-1$ minor obtained by deleting the ith row and j^{th} column of $\mathbf{A}$. The matrix of cofactors is denoted by $\mathring{\mathbf{A}}$.

Example 6.2.3

Problem: For $\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 6 \\ -3 & 9 & 1 \end{pmatrix}$, determine the cofactors $\mathring{A}_{21}$ and $\mathring{A}_{13}$.

Solution:

$\mathring{A}_{21} = (-1)^{2+1} M_{21} = (-1)(-19) = 19$ and $\mathring{A}_{13} = (-1)^{1+3} M_{13} = (+1)(18) = 18.$

The entire matrix of cofactors is $\mathring{\mathbf{A}} = \begin{pmatrix} -54 & -20 & 18 \\ 19 & 7 & -6 \\ -6 & -2 & 2 \end{pmatrix}$.

The cofactors of a square matrix $\mathbf{A}$ appear naturally in the expansion of $\det(\mathbf{A})$. For example,

$$\begin{aligned}\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\qquad - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) \\ &\qquad + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\ &= a_{11}\mathring{A}_{11} + a_{12}\mathring{A}_{12} + a_{13}\mathring{A}_{13}.\end{aligned} \tag{6.2.4}$$

Because this expansion is in terms of the entries of the first row and the corresponding cofactors, (6.2.4) is called *the cofactor expansion of* $\det(\mathbf{A})$ *in terms of the first row.* It should be clear that there is nothing special about the first row of $\mathbf{A}$. That is, it's just as easy to write an expression similar to (6.2.4) in which entries from any other row or column appear. For example, the terms in (6.2.4) can be rearranged to produce

$$\begin{aligned}\det(\mathbf{A}) &= a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{22}(a_{11}a_{33} - a_{13}a_{31}) + a_{32}(a_{13}a_{21} - a_{11}a_{23}) \\ &= a_{12}\mathring{A}_{12} + a_{22}\mathring{A}_{22} + a_{32}\mathring{A}_{32}.\end{aligned}$$

This is called *the cofactor expansion for* $\det(\mathbf{A})$ *in terms of the second column.* The 3×3 case is typical, and exactly the same reasoning can be applied to a more general $n \times n$ matrix in order to obtain the following statements.

Cofactor Expansions

- $\det(\mathbf{A}) = a_{i1}\mathring{A}_{i1} + a_{i2}\mathring{A}_{i2} + \cdots + a_{in}\mathring{A}_{in}$ (about row i). (6.2.5)
- $\det(\mathbf{A}) = a_{1j}\mathring{A}_{1j} + a_{2j}\mathring{A}_{2j} + \cdots + a_{nj}\mathring{A}_{nj}$ (about column j). (6.2.6)

Example 6.2.4

Problem: Use cofactor expansions to evaluate $\det(\mathbf{A})$ for

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

Solution: To minimize the effort, expand $\det(\mathbf{A})$ in terms of the row or column that contains a maximal number of zeros. For this example, the expansion in terms of the first row is most efficient because

$$\det(\mathbf{A}) = a_{11}\mathring{A}_{11} + a_{12}\mathring{A}_{12} + a_{13}\mathring{A}_{13} + a_{14}\mathring{A}_{14} = a_{14}\mathring{A}_{14} = (2)(-1)\begin{vmatrix} 7 & 1 & 6 \\ 3 & 7 & 2 \\ 0 & 3 & -1 \end{vmatrix}.$$

Now expand this remaining 3×3 determinant either in terms of the first column or the third row. Using the first column produces

$$\begin{vmatrix} 7 & 1 & 6 \\ 3 & 7 & 2 \\ 0 & 3 & -1 \end{vmatrix} = (7)(+1)\begin{vmatrix} 7 & 2 \\ 3 & -1 \end{vmatrix} + (3)(-1)\begin{vmatrix} 1 & 6 \\ 3 & -1 \end{vmatrix} = -91 + 57 = -34,$$

so $\det(\mathbf{A}) = (2)(-1)(-34) = 68$. You may wish to try an expansion using different rows or columns, and verify that the final result is the same.

In the previous example, we were able to take advantage of the fact that there were zeros in convenient positions. However, for a general matrix $\mathbf{A}_{n\times n}$ with no zero entries, it's not difficult to verify that successive application of cofactor expansions requires $n!\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(n-1)!}\right)$ multiplications to evaluate $\det(\mathbf{A})$. Even for moderate values of n, this number is too large for the cofactor expansion to be practical for computational purposes. Nevertheless, cofactors can be useful for theoretical developments such as the following determinant formula for $\mathbf{A}^{-1}$.

Determinant Formula for $\mathbf{A}^{-1}$

The ***adjugate*** of $\mathbf{A}_{n\times n}$ is defined to be $\operatorname{adj}(\mathbf{A}) = \mathring{\mathbf{A}}^T$, the transpose of the matrix of cofactors—some older texts call this the ***adjoint*** matrix. If $\mathbf{A}$ is nonsingular, then

$$\mathbf{A}^{-1} = \frac{\mathring{\mathbf{A}}^T}{\det(\mathbf{A})} = \frac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})}. \tag{6.2.7}$$

Proof. $\left[\mathbf{A}^{-1}\right]_{ij}$ is the ith component in the solution to $\mathbf{A}\mathbf{x} = \mathbf{e}_j$, where $\mathbf{e}_j$ is the jth unit vector. By Cramer's rule, this is

$$\left[\mathbf{A}^{-1}\right]_{ij} = x_i = \frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})},$$

where $\mathbf{A}_i$ is identical to $\mathbf{A}$ except that the ith column has been replaced by $\mathbf{e}_j$, and the cofactor expansion in terms of the ith column implies that

$$\det(\mathbf{A}_i) = \begin{array}{c} \begin{array}{ccccc} & & i^{th} & & \\ & & \downarrow & & \end{array} \\ \begin{vmatrix} a_{11} & \cdots & 0 & \cdots & a_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{j1} & \cdots & 1 & \cdots & a_{jn} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{n1} & \cdots & 0 & \cdots & a_{nn} \end{vmatrix} \end{array} = \mathring{A}_{ji}. \quad \blacksquare$$

Example 6.2.5

Problem: Use determinants to compute $\left[\mathbf{A}^{-1}\right]_{12}$ and $\left[\mathbf{A}^{-1}\right]_{31}$ for the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & 6 \\ -3 & 9 & 1 \end{pmatrix}.$$

Solution: The cofactors $\mathring{A}_{21}$ and $\mathring{A}_{13}$ were determined in Example 6.2.3 to be $\mathring{A}_{21} = 19$ and $\mathring{A}_{13} = 18$, and it's straightforward to compute $\det(\mathbf{A}) = 2$, so

$$\left[\mathbf{A}^{-1}\right]_{12} = \frac{\mathring{A}_{21}}{\det(\mathbf{A})} = \frac{19}{2} \quad \text{and} \quad \left[\mathbf{A}^{-1}\right]_{31} = \frac{\mathring{A}_{13}}{\det(\mathbf{A})} = \frac{18}{2} = 9.$$

Using the matrix of cofactors $\mathring{\mathbf{A}}$ computed in Example 6.2.3, we have that

$$\mathbf{A}^{-1} = \frac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})} = \frac{\mathring{\mathbf{A}}^T}{\det(\mathbf{A})} = \frac{1}{2}\begin{pmatrix} -54 & 19 & -6 \\ -20 & 7 & -2 \\ 18 & -6 & 2 \end{pmatrix}.$$

Example 6.2.6

Problem: For $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, determine a general formula for $\mathbf{A}^{-1}$.

Solution: $\operatorname{adj}(\mathbf{A}) = \mathbf{A}^T = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$, and $\det(\mathbf{A}) = ad - bc$, so

$$\mathbf{A}^{-1} = \frac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Example 6.2.7

Problem: Explain why the entries in $\mathbf{A}^{-1}$ vary continuously with the entries in $\mathbf{A}$ when $\mathbf{A}$ is nonsingular. This is in direct contrast with the lack of continuity exhibited by pseudoinverses (p. 423).

Solution: Recall from elementary calculus that the sum, the product, and the quotient of continuous functions are each continuous functions. In particular, the sum and the product of any set of numbers varies continuously as the numbers vary, so $\det(\mathbf{A})$ is a continuous function of the a_{ij}'s. Since each entry in $\operatorname{adj}(\mathbf{A})$ is a determinant, each quotient $[\mathbf{A}^{-1}]_{ij} = [\operatorname{adj}(\mathbf{A})]_{ij}/\det(\mathbf{A})$ must be a continuous function of the a_{ij}'s.

The Moral: The formula $\mathbf{A}^{-1} = \operatorname{adj}(\mathbf{A})/\det(\mathbf{A})$ is nearly worthless for actually computing the value of $\mathbf{A}^{-1}$, but, as this example demonstrates, the formula is nevertheless a useful mathematical tool. It's not uncommon for applied oriented students to fall into the trap of believing that the worth of a formula or an idea is tied to its utility for computing something. This example makes the point that things can have significant mathematical value without being computationally important. In fact, most of this chapter is in this category.

Example 6.2.8

Problem: Explain why the inner product of one row (or column) in $\mathbf{A}_{n\times n}$ with the cofactors of a different row (or column) in $\mathbf{A}$ must always be zero.

Solution: Let $\tilde{\mathbf{A}}$ be the result of replacing the jth column in $\mathbf{A}$ by the kth column of $\mathbf{A}$. Since $\tilde{\mathbf{A}}$ has two identical columns, $\det(\tilde{\mathbf{A}}) = 0$. Furthermore, the cofactor associated with the (i,j)-position in $\tilde{\mathbf{A}}$ is $\mathring{A}_{ij}$, the cofactor associated with the (i,j) in $\mathbf{A}$, so expansion of $\det(\tilde{\mathbf{A}})$ in terms of the jth column yields

$$0 = \det(\tilde{\mathbf{A}}) = \begin{array}{c} \begin{array}{ccccccc} & & j^{th} & & k^{th} & & \\ & & \downarrow & & \downarrow & & \end{array} \\ \begin{vmatrix} a_{11} & \cdots & a_{1k} & \cdots & a_{1k} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots & & \\ a_{i1} & \cdots & a_{ik} & \cdots & a_{ik} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots & & \\ a_{n1} & \cdots & a_{nk} & \cdots & a_{nk} & \cdots & a_{nn} \end{vmatrix} \end{array} = \sum_{i=1}^{n} a_{ik}\mathring{A}_{ij}.$$

Thus the inner product of the kth column of $\mathbf{A}_{n\times n}$ with the cofactors of the jth column of $\mathbf{A}$ is zero. A similar result holds for rows. Combining these observations with (6.2.5) and (6.2.6) produces

$$\sum_{j=1}^{n} a_{kj}\mathring{A}_{ij} = \begin{cases} \det(\mathbf{A}) & \text{if } k = i, \\ 0 & \text{if } k \neq i, \end{cases} \quad \text{and} \quad \sum_{i=1}^{n} a_{ik}\mathring{A}_{ij} = \begin{cases} \det(\mathbf{A}) & \text{if } k = j, \\ 0 & \text{if } k \neq j, \end{cases}$$

which is equivalent to saying that $\mathbf{A}[\operatorname{adj}(\mathbf{A})] = [\operatorname{adj}(\mathbf{A})]\mathbf{A} = \det(\mathbf{A})\,\mathbf{I}$.

Example 6.2.9

Differential Equations and Determinants. A system of n homogeneous first-order linear differential equations

$$\frac{d\,x_i(t)}{dt} = a_{i1}(t)x_1(t) + a_{i2}(t)x_2(t) + \cdots + a_{in}(t)x_n(t), \quad i = 1, 2, \ldots, n$$

can be expressed in matrix notation by writing

$$\begin{pmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_n'(t) \end{pmatrix} = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}$$

or, equivalently, $\mathbf{x}' = \mathbf{A}\mathbf{x}$. Let $\mathcal{S} = \{\mathbf{w}_1(t), \mathbf{w}_2(t), \ldots, \mathbf{w}_n(t)\}$ be a set of $n \times 1$ vectors that are solutions to $\mathbf{x}' = \mathbf{A}\mathbf{x}$, and place these solutions as columns in a matrix $\mathbf{W}(t)_{n\times n} = [\mathbf{w}_1(t) \,|\, \mathbf{w}_2(t) \,|\, \cdots \,|\, \mathbf{w}_n(t)]$ so that $\mathbf{W}' = \mathbf{A}\mathbf{W}$.

Problem: Prove that if $w(t) = \det(\mathbf{W})$, (called the *Wronskian* (p. 474)), then

$$w(t) = w(\xi_0)\, e^{\int_{\xi_0}^{t} \operatorname{trace} \mathbf{A}(\xi)\, d\xi}, \quad \text{where } \xi_0 \text{ is an arbitrary constant.} \tag{6.2.8}$$

Solution: By (6.1.19), $d\,w(t)/dt = \sum_{i=1}^{n} \det(\mathbf{D}_i)$, where

$$\mathbf{D}_i = \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ w'_{i1} & w'_{i2} & \cdots & w'_{in} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{pmatrix} = \mathbf{W} + \mathbf{e}_i\mathbf{e}_i^T\mathbf{W}' - \mathbf{e}_i\mathbf{e}_i^T\mathbf{W}.$$

Notice that $\left(-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}\right)$ subtracts $\mathbf{W}_{i*}$ from the ith row while $\left(+\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}'\right)$ adds $\mathbf{W}'_{i*}$ to the ith row. Use the fact that $\mathbf{W}' = \mathbf{A}\mathbf{W}$ to write

$$\mathbf{D}_i = \mathbf{W}+\mathbf{e}_i\mathbf{e}_i^T\mathbf{W}'-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W} = \mathbf{W}+\mathbf{e}_i\mathbf{e}_i^T\mathbf{A}\mathbf{W}-\mathbf{e}_i\mathbf{e}_i^T\mathbf{W} = \left(\mathbf{I}+\mathbf{e}_i\left(\mathbf{e}_i^T\mathbf{A}-\mathbf{e}_i^T\right)\right)\mathbf{W},$$

and apply formula (6.2.2) for the determinant of a rank-one updated matrix together with the product rule (6.1.15) to produce

$$\det(\mathbf{D}_i) = \left(1 + \mathbf{e}_i^T\mathbf{A}\mathbf{e}_i - \mathbf{e}_i^T\mathbf{e}_i\right)\det(\mathbf{W}) = a_{ii}(t)\,w(t),$$

so

$$\frac{d\,w(t)}{dt} = \sum_{i=1}^{n} \det(\mathbf{D}_i) = \left(\sum_{i=1}^{n} a_{ii}(t)\right) w(t) = trace\,\mathbf{A}(t)\,w(t).$$

In other words, $w(t)$ satisfies the first-order differential equation $w' = \tau\, w$, where $\tau = trace\,\mathbf{A}(t)$, and the solution of this equation is $w(t) = w(\xi_0)\,\mathrm{e}^{\int_{\xi_0}^{t} \tau(\xi)\,d\xi}$.

Consequences: In addition to its aesthetic elegance, (6.2.8) is a useful result because it is the basis for the following theorems.

- If $\mathbf{x}' = \mathbf{A}\mathbf{x}$ has a set of solutions $\mathcal{S} = \{\mathbf{w}_1(t), \mathbf{w}_2(t), \ldots, \mathbf{w}_n(t)\}$ that is linearly independent at *some* point $\xi_0 \in (a,b)$, and if $\int_{\xi_0}^{t} \tau(\xi)\,d\xi$ is finite for $t \in (a,b)$, then $\mathcal{S}$ must be linearly independent at *every* point $t \in (a,b)$.
- If $\mathbf{A}$ is a constant matrix, and if $\mathcal{S}$ is a set of n solutions that is linearly independent at *some* value $t = \xi_0$, then $\mathcal{S}$ must be linearly independent for *all* values of t.

Proof. If $\mathcal{S}$ is linearly independent at ξ_0, then $\mathbf{W}(\xi_0)$ is nonsingular, so $w(\xi_0) \neq 0$. If $\int_{\xi_0}^{t} \tau(\xi)\,d\xi$ is finite when $t \in (a,b)$, then $\mathrm{e}^{\int_{\xi_0}^{t} \tau(\xi)\,d\xi}$ is finite and nonzero on (a,b), so, by (6.2.8), $w(t) \neq 0$ on (a,b). Therefore, $\mathbf{W}(t)$ is nonsingular for $t \in (a,b)$, and thus $\mathcal{S}$ is linearly independent at each $t \in (a,b)$.

Exercises for section 6.2

6.2.1. Use a cofactor expansion to evaluate each of the following determinants.

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

6.2.2. Use determinants to compute the following inverses.

(a) $\begin{pmatrix} 2 & 1 & 1 \\ 6 & 2 & 1 \\ -2 & 2 & 1 \end{pmatrix}^{-1}$. (b) $\begin{pmatrix} 0 & 0 & -2 & 3 \\ 1 & 0 & 1 & 2 \\ -1 & 1 & 2 & 1 \\ 0 & 2 & -3 & 0 \end{pmatrix}^{-1}$.

6.2.3. (a) Use Cramer's rule to solve

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1 + x_2 &= \alpha, \\ x_2 + x_3 &= \beta. \end{aligned}$$

(b) Evaluate $\lim_{t\to\infty} x_2(t)$, where $x_2(t)$ is defined by the system

$$\begin{aligned} x_1 + tx_2 + t^2x_3 &= t^4, \\ t^2x_1 + x_2 + tx_3 &= t^3, \\ tx_1 + t^2x_2 + x_3 &= 0. \end{aligned}$$

6.2.4. Is the following equation a valid derivation of Cramer's rule for solving a nonsingular system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}_i$ is as described on p. 476?

$$\frac{\det(\mathbf{A}_i)}{\det(\mathbf{A})} = \det\left(\mathbf{A}^{-1}\mathbf{A}_i\right) = \det\left[\mathbf{e}_1 \cdots \mathbf{e}_{i-1}\ \mathbf{x}\ \mathbf{e}_{i+1} \cdots \mathbf{e}_n\right] = x_i.$$

6.2.5. (a) By example, show that $\det(\mathbf{A}+\mathbf{B}) \neq \det(\mathbf{A}) + \det(\mathbf{B})$.

(b) Using square matrices, construct an example that shows that

$$\det\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \neq \det(\mathbf{A})\det(\mathbf{D}) - \det(\mathbf{B})\det(\mathbf{C}).$$

6.2.6. Suppose $rank\,(\mathbf{B}_{m\times n}) = n$, and let $\mathbf{Q}$ be the orthogonal projector onto $N\left(\mathbf{B}^T\right)$. For $\mathbf{A} = [\mathbf{B}\,|\,\mathbf{c}_{n\times 1}]$, prove $\mathbf{c}^T\mathbf{Qc} = \det\left(\mathbf{A}^T\mathbf{A}\right)/\det\left(\mathbf{B}^T\mathbf{B}\right)$.

6.2.7. If $\mathbf{A}_{n\times n}$ is a nonsingular matrix, and if $\mathbf{D}$ and $\mathbf{C}$ are $n \times k$ matrices, explain how to use (6.2.1) to derive the formula

$$\det\left(\mathbf{A} + \mathbf{CD}^T\right) = \det(\mathbf{A})\det\left(\mathbf{I}_k + \mathbf{D}^T\mathbf{A}^{-1}\mathbf{C}\right).$$

Note: This is a generalization of (6.2.3) because if $\mathbf{c}_i$ and $\mathbf{d}_i$ are the ith columns of $\mathbf{C}$ and $\mathbf{D}$, respectively, then

$$\mathbf{A} + \mathbf{CD}^T = \mathbf{A} + \mathbf{c}_1\mathbf{d}_1^T + \mathbf{c}_2\mathbf{d}_2^T + \cdots + \mathbf{c}_k\mathbf{d}_k^T.$$

6.2.8. Explain why $\mathbf{A}$ is singular if and only if $\mathbf{A}[\text{adj}\,(\mathbf{A})] = \mathbf{0}$.

6.2.9. For a nonsingular linear system $\mathbf{Ax} = \mathbf{b}$, explain why each component of the solution must vary continuously with the entries of $\mathbf{A}$.

6.2.10. For scalars α, explain why $\text{adj}\,(\alpha\mathbf{A}) = \alpha^{n-1}\text{adj}\,(\mathbf{A})$. **Hint:** Recall Exercise 6.1.11.

6.2.11. For an $n \times n$ matrix $\mathbf{A}$, prove that the following statements are true.

(a) If $rank\,(\mathbf{A}) < n-1$, then $\text{adj}\,(\mathbf{A}) = \mathbf{0}$.

(b) If $rank\,(\mathbf{A}) = n-1$, then $rank\,(\text{adj}\,(\mathbf{A})) = 1$.

(c) If $rank\,(\mathbf{A}) = n$, then $rank\,(\text{adj}\,(\mathbf{A})) = n$.

6.2.12. In 1812, Cauchy discovered the formula that says that if $\mathbf{A}$ is $n \times n$, then $\det\,(\text{adj}\,(\mathbf{A})) = [\det\,(\mathbf{A})]^{n-1}$. Establish Cauchy's formula.

6.2.13. For the following tridiagonal matrix, $\mathbf{A}_n$, let $D_n = \det\,(\mathbf{A}_n)$, and derive the formula $D_n = 2D_{n-1} - D_{n-2}$ to deduce that $D_n = n+1$.

$$\mathbf{A}_n = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}_{n\times n}.$$

6.2.14. By considering rank-one updated matrices, derive the following formulas.

(a) $$\begin{vmatrix} \frac{1+\alpha_1}{\alpha_1} & 1 & \cdots & 1 \\ 1 & \frac{1+\alpha_2}{\alpha_2} & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & \frac{1+\alpha_n}{\alpha_n} \end{vmatrix} = \frac{1+\sum \alpha_i}{\prod \alpha_i}.$$

(b) $$\begin{vmatrix} \alpha & \beta & \beta & \cdots & \beta \\ \beta & \alpha & \beta & \cdots & \beta \\ \beta & \beta & \alpha & \cdots & \beta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \beta & \cdots & \alpha \end{vmatrix}_{n\times n} = \begin{cases} (\alpha-\beta)^n\left(1+\frac{n\beta}{\alpha-\beta}\right) & \text{if } \alpha \neq \beta, \\ 0 & \text{if } \alpha = \beta. \end{cases}$$

(c) $$\begin{vmatrix} 1+\alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1 & 1+\alpha_2 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & 1+\alpha_n \end{vmatrix} = 1+\alpha_1+\alpha_2+\cdots+\alpha_n.$$

6.2.15. A ***bordered matrix*** has the form $\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & \alpha \end{pmatrix}$ in which $\mathbf{A}_{n\times n}$ is nonsingular, $\mathbf{x}$ is a column, $\mathbf{y}^T$ is a row, and α is a scalar. Explain why the following statements must be true.

(a) $\begin{vmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & -1 \end{vmatrix} = -\det\left(\mathbf{A} + \mathbf{x}\mathbf{y}^T\right)$. (b) $\begin{vmatrix} \mathbf{A} & \mathbf{x} \\ \mathbf{y}^T & 0 \end{vmatrix} = -\mathbf{y}^T \operatorname{adj}(\mathbf{A})\,\mathbf{x}$.

6.2.16. If $\mathbf{B}$ is $m \times n$ and $\mathbf{C}$ is $n \times m$, explain why (6.2.1) guarantees that $\lambda^m \det(\lambda \mathbf{I}_n - \mathbf{CB}) = \lambda^n \det(\lambda \mathbf{I}_m - \mathbf{BC})$ is true for all scalars λ.

6.2.17. For a square matrix $\mathbf{A}$ and column vectors $\mathbf{c}$ and $\mathbf{d}$, derive the following two extensions of formula (6.2.3).

(a) If $\mathbf{Ax} = \mathbf{c}$, then $\det\left(\mathbf{A} + \mathbf{c}\mathbf{d}^T\right) = \det(\mathbf{A})\left(1 + \mathbf{d}^T\mathbf{x}\right)$.

(b) If $\mathbf{y}^T\mathbf{A} = \mathbf{d}^T$, then $\det\left(\mathbf{A} + \mathbf{c}\mathbf{d}^T\right) = \det(\mathbf{A})\left(1 + \mathbf{y}^T\mathbf{c}\right)$.

6.2.18. Describe the determinant of an elementary reflector (p. 324) and a plane rotation (p. 333), and then explain how to find $\det(\mathbf{A})$ using Householder reduction (p. 341) and Givens reduction (Example 5.7.2).

6.2.19. Suppose that $\mathbf{A}$ is a nonsingular matrix whose entries are integers. Prove that the entries in $\mathbf{A}^{-1}$ are integers if and only if $\det(\mathbf{A}) = \pm 1$.

6.2.20. Let $\mathbf{A} = \mathbf{I} - 2\mathbf{u}\mathbf{v}^T$ be a matrix in which $\mathbf{u}$ and $\mathbf{v}$ are column vectors with integer entries.

(a) Prove that $\mathbf{A}^{-1}$ has integer entries if and only if $\mathbf{v}^T\mathbf{u} = 0$ or 1.

(b) A matrix is said to be ***involutory*** whenever $\mathbf{A}^{-1} = \mathbf{A}$. Explain why $\mathbf{A} = \mathbf{I} - 2\mathbf{u}\mathbf{v}^T$ is involutory when $\mathbf{v}^T\mathbf{u} = 1$.

6.2.21. Use induction to argue that a cofactor expansion of $\det(\mathbf{A}_{n\times n})$ requires

$$c(n) = n!\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(n-1)!}\right)$$

multiplications for $n \geq 2$. Assume a computer will do 1,000,000 multiplications per second, and neglect all other operations to estimate how long it will take to evaluate the determinant of a 100×100 matrix using cofactor expansions. **Hint:** Recall the series expansion for e^x, and use $100! \approx 9.33 \times 10^{157}$.

6.2.22. Determine all values of λ for which the matrix $\mathbf{A}-\lambda\mathbf{I}$ is singular, where

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

Hint: If $p(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0$ is a monic polynomial with integer coefficients, then the integer roots of $p(\lambda)$ are a subset of the factors of α_0.

6.2.23. Suppose that $f_1(t),\ f_2(t), \ldots, f_n(t)$ are solutions of nth-order linear differential equation $y^{(n)} + p_1(t)y^{(n-1)} + \cdots + p_{n-1}(t)y' + p_n(t)y = 0$, and let $w(t)$ be the Wronskian

$$w(t) = \begin{vmatrix} f_1(t) & f_2(t) & \cdots & f_n(t) \\ f_1'(t) & f_2'(t) & \cdots & f_n'(t) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(t) & f_2^{(n-1)}(t) & \cdots & f_n^{(n-1)}(t) \end{vmatrix}.$$

By converting the nth-order equation into a system of n first-order equations with the substitutions $x_1 = y,\ x_2 = y', \ldots, x_n = y^{(n-1)}$, show that $w(t) = w(\xi_0)\,\mathrm{e}^{-\int_{\xi_0}^{t} p_1(\xi)\,d\xi}$ for an arbitrary constant ξ_0.

6.2.24. Evaluate the ***Vandermonde determinant*** by showing

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{j>i}(x_j - x_i).$$

When is this nonzero (compare with Example 4.3.4)? **Hint:** For the polynomial $p(\lambda) = \begin{vmatrix} 1 & \lambda & \lambda^2 & \cdots & \lambda^{k-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{k-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^{k-1} \end{vmatrix}_{k\times k}$, use induction to find the degree of $p(\lambda)$, the roots of $p(\lambda)$, and the coefficient of λ^{k-1} in $p(\lambda)$.

6.2.25. Suppose that each entry in $\mathbf{A}_{n\times n} = [a_{ij}(x)]$ is a differentiable function of a real variable x. Use formula (6.1.19) to derive the formula

$$\frac{d\big(\det(\mathbf{A})\big)}{dx} = \sum_{j=1}^{n}\sum_{i=1}^{n} \frac{d\,a_{ij}}{dx}\,\mathring{A}_{ij}.$$

6.2.26. Consider the entries of $\mathbf{A}$ to be independent variables, and use formula (6.1.19) to derive the formula

$$\frac{\partial \det(\mathbf{A})}{\partial a_{ij}} = \mathring{A}_{ij}.$$

6.2.27. Laplace's Expansion. In 1772, the French mathematician Pierre-Simon Laplace (1749–1827) presented the following generalized version of the cofactor expansion. For an $n \times n$ matrix $\mathbf{A}$, let

$$\mathbf{A}(i_1 i_2 \cdots i_k \,|\, j_1 j_2 \cdots j_k) = \text{the } k \times k \text{ submatrix of } \mathbf{A} \text{ that lies on the intersection of rows } i_1, i_2, \ldots, i_k \text{ with columns } j_1, j_2, \ldots, j_k,$$

and let

$$M(i_1 i_2 \cdots i_k \,|\, j_1 j_2 \cdots j_k) = \text{the } n-k \times n-k \text{ minor determinant obtained by deleting rows } i_1, i_2, \ldots, i_k \text{ and columns } j_1, j_2, \ldots, j_k \text{ from } \mathbf{A}.$$

The ***cofactor*** of $\mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k)$ is defined to be the signed minor

$$\mathring{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) = (-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_k} M(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

This is consistent with the definition of cofactor given earlier because if $\mathbf{A}(i \,|\, j) = a_{ij}$, then $\mathring{A}(i \,|\, j) = (-1)^{i+j} M(i \,|\, j) = (-1)^{i+j} M_{ij} = \mathring{A}_{ij}$. For each fixed set of row indices $1 \le i_1 < \cdots < i_k \le n$,

$$\det(\mathbf{A}) = \sum_{1 \le j_1 < \cdots < j_k \le n} \det \mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) \mathring{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

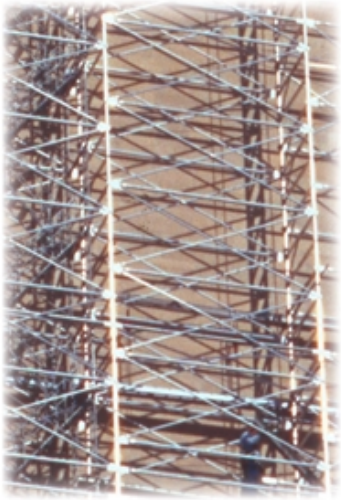
Similarly, for each fixed set of column indices $1 \le j_1 < \cdots < j_k \le n$,

$$\det(\mathbf{A}) = \sum_{1 \le i_1 < \cdots < i_k \le n} \det \mathbf{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k) \mathring{A}(i_1 \cdots i_k \,|\, j_1 \cdots j_k).$$

Each of these sums contains $\binom{n}{k}$ terms. Use Laplace's expansion to evaluate the determinant of

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

in terms of the first and third rows.



CHAPTER 7

Eigenvalues and Eigenvectors

7.1 ELEMENTARY PROPERTIES OF EIGENSYSTEMS

Up to this point, almost everything was either motivated by or evolved from the consideration of systems of linear *algebraic* equations. But we have come to a turning point, and from now on the emphasis will be different. Rather than being concerned with systems of *algebraic* equations, many topics will be motivated or driven by applications involving systems of linear *differential* equations and their discrete counterparts, difference equations.

For example, consider the problem of solving the system of two first-order linear differential equations, $du_1/dt = 7u_1 - 4u_2$ and $du_2/dt = 5u_1 - 2u_2$. In matrix notation, this system is

$$\begin{pmatrix} u_1' \\ u_2' \end{pmatrix} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{or, equivalently,} \quad \mathbf{u}' = \mathbf{A}\mathbf{u}, \tag{7.1.1}$$

where $\mathbf{u}' = \begin{pmatrix} u_1' \\ u_2' \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix}$, and $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$. Because solutions of a single equation $u' = \lambda u$ have the form $u = \alpha e^{\lambda t}$, we are motivated to seek solutions of (7.1.1) that also have the form

$$u_1 = \alpha_1 e^{\lambda t} \quad \text{and} \quad u_2 = \alpha_2 e^{\lambda t}. \tag{7.1.2}$$

Differentiating these two expressions and substituting the results in (7.1.1) yields

$$\begin{matrix} \alpha_1 \lambda e^{\lambda t} = 7\alpha_1 e^{\lambda t} - 4\alpha_2 e^{\lambda t} \\ \alpha_2 \lambda e^{\lambda t} = 5\alpha_1 e^{\lambda t} - 2\alpha_2 e^{\lambda t} \end{matrix} \Rightarrow \begin{matrix} \alpha_1 \lambda = 7\alpha_1 - 4\alpha_2 \\ \alpha_2 \lambda = 5\alpha_1 - 2\alpha_2 \end{matrix} \Rightarrow \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \lambda \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

In other words, solutions of (7.1.1) having the form (7.1.2) can be constructed provided solutions for λ and $\mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ in the matrix equation $\mathbf{Ax} = \lambda\mathbf{x}$ can be found. Clearly, $\mathbf{x} = \mathbf{0}$ trivially satisfies $\mathbf{Ax} = \lambda\mathbf{x}$, but $\mathbf{x} = \mathbf{0}$ provides no useful information concerning the solution of (7.1.1). What we really need are scalars λ and *nonzero* vectors $\mathbf{x}$ that satisfy $\mathbf{Ax} = \lambda\mathbf{x}$. Writing $\mathbf{Ax} = \lambda\mathbf{x}$ as $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$ shows that the vectors of interest are the *nonzero* vectors in $N(\mathbf{A} - \lambda\mathbf{I})$. But $N(\mathbf{A} - \lambda\mathbf{I})$ contains nonzero vectors if and only if $\mathbf{A} - \lambda\mathbf{I}$ is singular. Therefore, the scalars of interest are precisely the values of λ that make $\mathbf{A} - \lambda\mathbf{I}$ singular or, equivalently, the λ's for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. These observations motivate the definition of eigenvalues and eigenvectors.[66]

Eigenvalues and Eigenvectors

For an $n \times n$ matrix $\mathbf{A}$, scalars λ and vectors $\mathbf{x}_{n\times 1} \neq \mathbf{0}$ satisfying $\mathbf{Ax} = \lambda\mathbf{x}$ are called ***eigenvalues*** and ***eigenvectors*** of $\mathbf{A}$, respectively, and any such pair, $(\lambda, \mathbf{x})$, is called an ***eigenpair*** for $\mathbf{A}$. The set of *distinct* eigenvalues, denoted by $\sigma(\mathbf{A})$, is called the ***spectrum*** of $\mathbf{A}$.

- $\lambda \in \sigma(\mathbf{A}) \iff \mathbf{A} - \lambda\mathbf{I}$ is singular $\iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0.$ (7.1.3)
- $\{\mathbf{x} \neq \mathbf{0} \mid \mathbf{x} \in N(\mathbf{A} - \lambda\mathbf{I})\}$ is the set of all eigenvectors associated with λ. From now on, $N(\mathbf{A} - \lambda\mathbf{I})$ is called an ***eigenspace*** for $\mathbf{A}$.
- Nonzero row vectors $\mathbf{y}^*$ such that $\mathbf{y}^*(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$ are called ***left-hand eigenvectors*** for $\mathbf{A}$ (see Exercise 7.1.18 on p. 503).

Geometrically, $\mathbf{Ax} = \lambda\mathbf{x}$ says that under transformation by $\mathbf{A}$, eigenvectors experience only changes in magnitude or sign—the orientation of $\mathbf{Ax}$ in $\Re^n$ is the same as that of $\mathbf{x}$. The eigenvalue λ is simply the amount of "stretch" or "shrink" to which the eigenvector $\mathbf{x}$ is subjected when transformed by $\mathbf{A}$. Figure 7.1.1 depicts the situation in $\Re^2$.

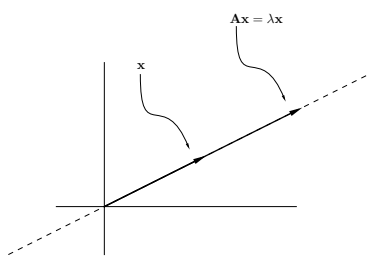


FIGURE 7.1.1

[66] The words *eigenvalue* and *eigenvector* are derived from the German word *eigen*, which means *owned by* or *peculiar to*. Eigenvalues and eigenvectors are sometimes called *characteristic values* and *characteristic vectors, proper values* and *proper vectors,* or *latent values* and *latent vectors.*

Let's now face the problem of finding the eigenvalues and eigenvectors of the matrix $\mathbf{A} = \begin{pmatrix} 7 & -4 \\ 5 & -2 \end{pmatrix}$ appearing in (7.1.1). As noted in (7.1.3), the eigenvalues are the scalars λ for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Expansion of $\det(\mathbf{A} - \lambda\mathbf{I})$ produces the second-degree polynomial

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 7-\lambda & -4 \\ 5 & -2-\lambda \end{vmatrix} = \lambda^2 - 5\lambda + 6 = (\lambda - 2)(\lambda - 3),$$

which is called the ***characteristic polynomial*** for $\mathbf{A}$. Consequently, the eigenvalues for $\mathbf{A}$ are the solutions of the ***characteristic equation*** $p(\lambda) = 0$ (i.e., the roots of the characteristic polynomial), and they are $\lambda = 2$ and $\lambda = 3$.

The eigenvectors associated with $\lambda = 2$ and $\lambda = 3$ are simply the nonzero vectors in the eigenspaces $N(\mathbf{A} - 2\mathbf{I})$ and $N(\mathbf{A} - 3\mathbf{I})$, respectively. But determining these eigenspaces amounts to nothing more than solving the two homogeneous systems, $(\mathbf{A} - 2\mathbf{I})\,\mathbf{x} = \mathbf{0}$ and $(\mathbf{A} - 3\mathbf{I})\,\mathbf{x} = \mathbf{0}$.

For $\lambda = 2$,

$$\mathbf{A} - 2\mathbf{I} = \begin{pmatrix} 5 & -4 \\ 5 & -4 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -4/5 \\ 0 & 0 \end{pmatrix} \implies \begin{array}{l} x_1 = (4/5)x_2 \\ x_2 \text{ is free} \end{array}$$

$$\implies N(\mathbf{A} - 2\mathbf{I}) = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \alpha \begin{pmatrix} 4/5 \\ 1 \end{pmatrix} \right\}.$$

For $\lambda = 3$,

$$\mathbf{A} - 3\mathbf{I} = \begin{pmatrix} 4 & -4 \\ 5 & -5 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 \\ 0 & 0 \end{pmatrix} \implies \begin{array}{l} x_1 = x_2 \\ x_2 \text{ is free} \end{array}$$

$$\implies N(\mathbf{A} - 3\mathbf{I}) = \left\{ \mathbf{x} \;\middle|\; \mathbf{x} = \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

In other words, the eigenvectors of $\mathbf{A}$ associated with $\lambda = 2$ are all nonzero multiples of $\mathbf{x} = (\,4/5 \quad 1\,)^T$, and the eigenvectors associated with $\lambda = 3$ are all nonzero multiples of $\mathbf{y} = (\,1 \quad 1\,)^T$. Although there are an infinite number of eigenvectors associated with each eigenvalue, each eigenspace is one dimensional, so, for this example, there is only one *independent* eigenvector associated with each eigenvalue.

Let's complete the discussion concerning the system of differential equations $\mathbf{u}' = \mathbf{A}\mathbf{u}$ in (7.1.1). Coupling (7.1.2) with the eigenpairs $(\lambda_1, \mathbf{x})$ and $(\lambda_2, \mathbf{y})$ of $\mathbf{A}$ computed above produces two solutions of $\mathbf{u}' = \mathbf{A}\mathbf{u}$, namely,

$$\mathbf{u}_1 = e^{\lambda_1 t}\mathbf{x} = e^{2t}\begin{pmatrix} 4/5 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = e^{\lambda_2 t}\mathbf{y} = e^{3t}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

It turns out that all other solutions are linear combinations of these two particular solutions—more is said in §7.4 on p. 541.

Below is a summary of some general statements concerning features of the characteristic polynomial and the characteristic equation.

Characteristic Polynomial and Equation

- The ***characteristic polynomial*** of $\mathbf{A}_{n\times n}$ is $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$. The degree of $p(\lambda)$ is n, and the leading term in $p(\lambda)$ is $(-1)^n\lambda^n$.
- The ***characteristic equation*** for $\mathbf{A}$ is $p(\lambda) = 0$.
- The eigenvalues of $\mathbf{A}$ are the solutions of the characteristic equation or, equivalently, the roots of the characteristic polynomial.
- Altogether, $\mathbf{A}$ has n eigenvalues, but some may be complex numbers (even if the entries of $\mathbf{A}$ are real numbers), and some eigenvalues may be repeated.
- If $\mathbf{A}$ contains only real numbers, then its complex eigenvalues must occur in conjugate pairs—i.e., if $\lambda \in \sigma(\mathbf{A})$, then $\overline{\lambda} \in \sigma(\mathbf{A})$.

Proof. The fact that $\det(\mathbf{A} - \lambda\mathbf{I})$ is a polynomial of degree n whose leading term is $(-1)^n\lambda^n$ follows from the definition of determinant given in (6.1.1). If

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

then

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \sum_p \sigma(p)(a_{1p_1} - \delta_{1p_1}\lambda)(a_{2p_2} - \delta_{2p_2}\lambda)\cdots(a_{np_n} - \delta_{np_n}\lambda)$$

is a polynomial in λ. The highest power of λ is produced by the term

$$(a_{11} - \lambda)(a_{22} - \lambda)\cdots(a_{nn} - \lambda),$$

so the degree is n, and the leading term is $(-1)^n\lambda^n$. The discussion given earlier contained the proof that the eigenvalues are precisely the solutions of the characteristic equation, but, for the sake of completeness, it's repeated below:

$$\begin{aligned} \lambda \in \sigma(\mathbf{A}) &\iff \mathbf{A}\mathbf{x} = \lambda\mathbf{x} \text{ for some } \mathbf{x} \neq \mathbf{0} \iff (\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0} \text{ for some } \mathbf{x} \neq \mathbf{0} \\ &\iff \mathbf{A} - \lambda\mathbf{I} \text{ is singular} \iff \det(\mathbf{A} - \lambda\mathbf{I}) = 0. \end{aligned}$$

The fundamental theorem of algebra is a deep result that insures every polynomial of degree n with real or complex coefficients has n roots, but some roots may be complex numbers (even if all the coefficients are real), and some roots may be repeated. Consequently, $\mathbf{A}$ has n eigenvalues, but some may be complex, and some may be repeated. The fact that complex eigenvalues of real matrices must occur in conjugate pairs is a consequence of the fact that the roots of a polynomial with real coefficients occur in conjugate pairs. ∎

Example 7.1.1

Problem: Determine the eigenvalues and eigenvectors of $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$.

Solution: The characteristic polynomial is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & -1 \\ 1 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 + 1 = \lambda^2 - 2\lambda + 2,$$

so the characteristic equation is $\lambda^2 - 2\lambda + 2 = 0$. Application of the quadratic formula yields

$$\lambda = \frac{2 \pm \sqrt{-4}}{2} = \frac{2 \pm 2\sqrt{-1}}{2} = 1 \pm \mathrm{i},$$

so the spectrum of $\mathbf{A}$ is $\sigma(\mathbf{A}) = \{1+\mathrm{i},\ 1-\mathrm{i}\}$. Notice that the eigenvalues are complex conjugates of each other—as they must be because complex eigenvalues of real matrices must occur in conjugate pairs. Now find the eigenspaces.

For $\lambda = 1 + \mathrm{i}$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} -\mathrm{i} & -1 \\ 1 & -\mathrm{i} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -\mathrm{i} \\ 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - \lambda\mathbf{I}) = span\left\{\begin{pmatrix} \mathrm{i} \\ 1 \end{pmatrix}\right\}.$$

For $\lambda = 1 - \mathrm{i}$,

$$\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} \mathrm{i} & -1 \\ 1 & \mathrm{i} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & \mathrm{i} \\ 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - \lambda\mathbf{I}) = span\left\{\begin{pmatrix} -\mathrm{i} \\ 1 \end{pmatrix}\right\}.$$

In other words, the eigenvectors associated with $\lambda_1 = 1 + \mathrm{i}$ are all nonzero multiples of $\mathbf{x}_1 = \begin{pmatrix} \mathrm{i} & 1 \end{pmatrix}^T$, and the eigenvectors associated with $\lambda_2 = 1 - \mathrm{i}$ are all nonzero multiples of $\mathbf{x}_2 = \begin{pmatrix} -\mathrm{i} & 1 \end{pmatrix}^T$. In previous sections, you could be successful by thinking only in terms of real numbers and by dancing around those statements and issues involving complex numbers. But this example makes it clear that avoiding complex numbers, even when dealing with real matrices, is no longer possible—very innocent looking matrices, such as the one in this example, can possess complex eigenvalues and eigenvectors.

As we have seen, computing eigenvalues boils down to solving a polynomial equation. But determining solutions to polynomial equations can be a formidable task. It was proven in the nineteenth century that it's impossible to express the roots of a general polynomial of degree five or higher using radicals of the coefficients. This means that there does not exist a generalized version of the quadratic formula for polynomials of degree greater than four, and general polynomial equations cannot be solved by a finite number of arithmetic operations involving $+, -, \times, \div, \sqrt[n]{\ }$. Unlike solving $\mathbf{Ax} = \mathbf{b}$, the eigenvalue problem generally requires an infinite algorithm, so all practical eigenvalue computations are accomplished by iterative methods—some are discussed later.

For theoretical work, and for textbook-type problems, it's helpful to express the characteristic equation in terms of the principal minors. Recall that an $r \times r$ ***principal submatrix*** of $\mathbf{A}_{n\times n}$ is a submatrix that lies on the same set of r rows and columns, and an $r \times r$ ***principal minor*** is the determinant of an $r \times r$ principal submatrix. In other words, $r \times r$ principal minors are obtained by deleting the same set of $n-r$ rows and columns, and there are $\binom{n}{r} = n!/r!(n-r)!$ such minors. For example, the 1×1 principal minors of

$$\mathbf{A} = \begin{pmatrix} -3 & 1 & -3 \\ 20 & 3 & 10 \\ 2 & -2 & 4 \end{pmatrix} \tag{7.1.4}$$

are the diagonal entries -3, 3, and 4. The 2×2 principal minors are

$$\begin{vmatrix} -3 & 1 \\ 20 & 3 \end{vmatrix} = -29, \quad \begin{vmatrix} -3 & -3 \\ 2 & 4 \end{vmatrix} = -6, \quad \text{and} \quad \begin{vmatrix} 3 & 10 \\ -2 & 4 \end{vmatrix} = 32,$$

and the only 3×3 principal minor is $\det(\mathbf{A}) = -18$.

Related to the principal minors are the symmetric functions of the eigenvalues. The kth ***symmetric function*** of $\lambda_1, \lambda_2, \ldots, \lambda_n$ is defined to be the sum of the product of the eigenvalues taken k at a time. That is,

$$s_k = \sum_{1 \le i_1 < \cdots < i_k \le n} \lambda_{i_1} \cdots \lambda_{i_k}.$$

For example, when $n = 4$,

$$\begin{aligned} s_1 &= \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4, \\ s_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_3\lambda_4, \\ s_3 &= \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_3\lambda_4 + \lambda_2\lambda_3\lambda_4, \\ s_4 &= \lambda_1\lambda_2\lambda_3\lambda_4. \end{aligned}$$

The connection between symmetric functions, principal minors, and the coefficients in the characteristic polynomial is given in the following theorem.

Coefficients in the Characteristic Equation

If $\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_{n-1}\lambda + c_n = 0$ is the characteristic equation for $\mathbf{A}_{n\times n}$, and if s_k is the kth symmetric function of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of $\mathbf{A}$, then

- $c_k = (-1)^k \sum(\text{all } k \times k \text{ principal minors}),$ (7.1.5)
- $s_k = \sum(\text{all } k \times k \text{ principal minors}),$ (7.1.6)
- $trace\,(\mathbf{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_n = -c_1,$ (7.1.7)
- $\det(\mathbf{A}) = \lambda_1\lambda_2 \cdots \lambda_n = (-1)^n c_n.$ (7.1.8)

Proof. At least two proofs of (7.1.5) are possible, and although they are conceptually straightforward, each is somewhat tedious. One approach is to successively use the result of Exercise 6.1.14 to expand $\det(\mathbf{A} - \lambda\mathbf{I})$. Another proof rests on the observation that if

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = (-1)^n\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n$$

is the characteristic *polynomial* for $\mathbf{A}$, then the characteristic *equation* is

$$\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_{n-1}\lambda + c_n = 0, \qquad \text{where} \qquad c_i = (-1)^n a_i.$$

Taking the rth derivative of $p(\lambda)$ yields $p^{(r)}(0) = r!a_{n-r}$, and hence

$$c_{n-r} = \frac{(-1)^n}{r!}p^{(r)}(0). \tag{7.1.9}$$

It's now a matter of repeatedly applying the formula (6.1.19) for differentiating a determinant to $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$. After r applications of (6.1.19),

$$p^{(r)}(\lambda) = \sum_{i_j \neq i_k} D_{i_1\cdots i_r}(\lambda),$$

where $D_{i_1\cdots i_r}(\lambda)$ is the determinant of the matrix identical to $\mathbf{A} - \lambda\mathbf{I}$ except that rows $i_1, i_2, \ldots, i_r$ have been replaced by $-\mathbf{e}_{i_1}^T, -\mathbf{e}_{i_2}^T, \ldots, -\mathbf{e}_{i_r}^T$, respectively. It follows that $D_{i_1\cdots i_r}(0) = (-1)^r\det(\mathbf{A}_{i_1\cdots i_r})$, where $\mathbf{A}_{i_1 i_2\cdots i_r}$ is identical to $\mathbf{A}$ except that rows $i_1, i_2, \ldots, i_r$ have been replaced by $\mathbf{e}_{i_1}^T, \mathbf{e}_{i_2}^T, \ldots, \mathbf{e}_{i_r}^T$, respectively, and $\det(\mathbf{A}_{i_1\cdots i_r})$ is the $n-r \times n-r$ principal minor obtained by deleting rows and columns $i_1, i_2, \ldots, i_r$ from $\mathbf{A}$. Consequently,

$$\begin{aligned} p^{(r)}(0) &= \sum_{i_j \neq i_k} D_{i_1\cdots i_r}(0) = (-1)^r \sum_{i_j \neq i_k} \det(\mathbf{A}_{i_1\cdots i_r}) \\ &= r! \times (-1)^r \sum(\text{all } n-r \times n-r \text{ principal minors}). \end{aligned}$$

The factor $r!$ appears because each of the $r!$ permutations of the subscripts on $\mathbf{A}_{i_1\cdots i_r}$ describes the same matrix. Therefore, (7.1.9) says

$$c_{n-r} = \frac{(-1)^n}{r!}p^{(r)}(0) = (-1)^{n-r}\sum(\text{all } n-r \times n-r \text{ principal minors}).$$

To prove (7.1.6), write the characteristic equation for $\mathbf{A}$ as

$$(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) = 0, \tag{7.1.10}$$

and expand the left-hand side to produce

$$\lambda^n - s_1\lambda^{n-1} + \cdots + (-1)^k s_k\lambda^{n-k} + \cdots + (-1)^n s_n = 0. \tag{7.1.11}$$

(Using $n = 3$ or $n = 4$ in (7.1.10) makes this clear.) Comparing (7.1.11) with (7.1.5) produces the desired conclusion. Statements (7.1.7) and (7.1.8) are obtained from (7.1.5) and (7.1.6) by setting $k = 1$ and $k = n$. ∎

Example 7.1.2

Problem: Determine the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} -3 & 1 & -3 \\ 20 & 3 & 10 \\ 2 & -2 & 4 \end{pmatrix}.$$

Solution: Use the principal minors computed in (7.1.4) along with (7.1.5) to obtain the characteristic equation

$$\lambda^3 - 4\lambda^2 - 3\lambda + 18 = 0.$$

A result from elementary algebra states that if the coefficients α_i in

$$\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 = 0$$

are integers, then every integer solution is a factor of α_0. For our problem, this means that if there exist integer eigenvalues, then they must be contained in the set $\mathcal{S} = \{\pm 1, \pm 2, \pm 3, \pm 6, \pm 9, \pm 18\}$. Evaluating $p(\lambda)$ for each $\lambda \in \mathcal{S}$ reveals that $p(3) = 0$ and $p(-2) = 0$, so $\lambda = 3$ and $\lambda = -2$ are eigenvalues for $\mathbf{A}$. To determine the other eigenvalue, deflate the problem by dividing

$$\frac{\lambda^3 - 4\lambda^2 - 3\lambda + 18}{\lambda - 3} = \lambda^2 - \lambda - 6 = (\lambda - 3)(\lambda + 2).$$

Thus the characteristic equation can be written in factored form as

$$(\lambda - 3)^2(\lambda + 2) = 0,$$

so the spectrum of $\mathbf{A}$ is $\sigma(\mathbf{A}) = \{3, -2\}$ in which $\lambda = 3$ is repeated—we say that the ***algebraic multiplicity*** of $\lambda = 3$ is two. The eigenspaces are obtained as follows.

For $\lambda = 3$,

$$\mathbf{A} - 3\mathbf{I} \longrightarrow \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies N(\mathbf{A} - 3\mathbf{I}) = span\left\{\begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}\right\}.$$

For $\lambda = -2$,

$$\mathbf{A} + 2\mathbf{I} \longrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{pmatrix} \implies N(\mathbf{A} + 2\mathbf{I}) = span\left\{\begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}\right\}.$$

Notice that although the algebraic multiplicity of $\lambda = 3$ is two, the dimension of the associated eigenspace is only one—we say that $\mathbf{A}$ is ***deficient*** in eigenvectors. As we will see later, deficient matrices pose significant difficulties.

Example 7.1.3

Continuity of Eigenvalues. A classical result (requiring complex analysis) states that the roots of a polynomial vary continuously with the coefficients. Since the coefficients of the characteristic polynomial $p(\lambda)$ of $\mathbf{A}$ can be expressed in terms of sums of principal minors, it follows that the coefficients of $p(\lambda)$ vary continuously with the entries of $\mathbf{A}$. Consequently, the eigenvalues of $\mathbf{A}$ must vary continuously with the entries of $\mathbf{A}$. **Caution!** Components of an eigenvector need not vary continuously with the entries of $\mathbf{A}$—e.g., consider $\mathbf{x} = (\epsilon^{-1}, 1)$ as an eigenvector for $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & \epsilon \end{pmatrix}$, and let $\epsilon \to 0$.

Example 7.1.4

Spectral Radius. For square matrices $\mathbf{A}$, the number

$$\rho(\mathbf{A}) = \max_{\lambda \in \sigma(\mathbf{A})} |\lambda|$$

is called the ***spectral radius*** of $\mathbf{A}$. It's not uncommon for applications to require only a bound on the eigenvalues of $\mathbf{A}$. That is, precise knowledge of each eigenvalue may not called for, but rather just an upper bound on $\rho(\mathbf{A})$ is all that's often needed. A rather crude (but cheap) upper bound on $\rho(\mathbf{A})$ is obtained by observing that $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$ for every matrix norm. This is true because if $(\lambda, \mathbf{x})$ is any eigenpair, then $\mathbf{X} = \big[\mathbf{x} \,|\, \mathbf{0} \,|\, \cdots \,|\, \mathbf{0}\big]_{n \times n} \neq \mathbf{0}$, and $\lambda \mathbf{X} = \mathbf{A}\mathbf{X}$ implies $|\lambda| \, \|\mathbf{X}\| = \|\lambda \mathbf{X}\| = \|\mathbf{A}\mathbf{X}\| \leq \|\mathbf{A}\| \, \|\mathbf{X}\|$, so

$$|\lambda| \leq \|\mathbf{A}\| \quad \text{for all } \lambda \in \sigma(\mathbf{A}). \tag{7.1.12}$$

This result is a precursor to a stronger relationship between spectral radius and norm that is hinted at in Exercise 7.3.12 and developed in Example 7.10.1 (p. 619).

The eigenvalue bound (7.1.12) given in Example 7.1.4 is cheap to compute, especially if the 1-norm or ∞-norm is used, but you often get what you pay for. You get one big circle whose radius is usually much larger than the spectral radius $\rho(\mathbf{A})$. It's possible to do better by using a set of Gerschgorin[67] circles as described below.

[67] S. A. Gerschgorin illustrated the use of Gerschgorin circles for estimating eigenvalues in 1931, but the concept appears earlier in work by L. Lévy in 1881, by H. Minkowski (p. 278) in 1900, and by J. Hadamard (p. 469) in 1903. However, each time the idea surfaced, it gained little attention and was quickly forgotten until Olga Taussky (1906–1995), the premier woman of linear algebra, and her fellow German emigrè Alfred Brauer (1894–1985) became captivated by the result. Taussky (who became Olga Taussky-Todd after marrying the numerical analyst John Todd) and Brauer devoted significant effort to strengthening, promoting, and popularizing Gerschgorin-type eigenvalue bounds. Their work during the 1940s and 1950s ended the periodic rediscoveries, and they made Gerschgorin (who might otherwise have been forgotten) famous.

Gerschgorin Circles

- The eigenvalues of $\mathbf{A} \in \mathcal{C}^{n\times n}$ are contained the union $\mathcal{G}_r$ of the n ***Gerschgorin circles*** defined by

$$|z - a_{ii}| \le r_i, \quad \text{where} \quad r_i = \sum_{\substack{j=1\\ j\ne i}}^{n} |a_{ij}| \text{ for } i = 1, 2, \ldots, n. \tag{7.1.13}$$

  In other words, the eigenvalues are trapped in the collection of circles centered at a_{ii} with radii given by the sum of absolute values in $\mathbf{A}_{i*}$ with a_{ii} deleted.

- Furthermore, if a union $\mathcal{U}$ of k Gerschgorin circles does not touch any of the other $n-k$ circles, then there are exactly k eigenvalues (counting multiplicities) in the circles in $\mathcal{U}$. (7.1.14)

- Since $\sigma(\mathbf{A}^T) = \sigma(\mathbf{A})$, the deleted absolute row sums in (7.1.13) can be replaced by deleted absolute column sums, so the eigenvalues of $\mathbf{A}$ are also contained in the union $\mathcal{G}_c$ of the circles defined by

$$|z - a_{jj}| \le c_j, \quad \text{where} \quad c_j = \sum_{\substack{i=1\\ i\ne j}}^{n} |a_{ij}| \text{ for } j = 1, 2, \ldots, n. \tag{7.1.15}$$

- Combining (7.1.13) and (7.1.15) means that the eigenvalues of $\mathbf{A}$ are contained in the intersection $\mathcal{G}_r \cap \mathcal{G}_c$. (7.1.16)

Proof. Let $(\lambda, \mathbf{x})$ be an eigenpair for $\mathbf{A}$, and assume $\mathbf{x}$ has been normalized so that $\|\mathbf{x}\|_\infty = 1$. If x_i is a component of $\mathbf{x}$ such that $|x_i| = 1$, then

$$\lambda x_i = [\lambda \mathbf{x}]_i = [\mathbf{A}\mathbf{x}]_i = \sum_{j=1}^{n} a_{ij}x_j \quad \Longrightarrow \quad (\lambda - a_{ii})x_i = \sum_{\substack{j=1\\ j\ne i}}^{n} a_{ij}x_j,$$

and hence

$$|\lambda - a_{ii}| = |\lambda - a_{ii}|\,|x_i| = \left|\sum_{j\ne i} a_{ij}x_j\right| \le \sum_{j\ne i} |a_{ij}|\,|x_j| \le \sum_{j\ne i} |a_{ij}| = r_i.$$

Thus λ is in one of the Gerschgorin circles, so the union of all such circles contains $\sigma(\mathbf{A})$. To establish (7.1.14), let $\mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ and $\mathbf{B} = \mathbf{A} - \mathbf{D}$, and set $\mathbf{C}(t) = \mathbf{D} + t\mathbf{B}$ for $t \in [0, 1]$. The first part shows that the eigenvalues of $\lambda_i(t)$ of $\mathbf{C}(t)$ are contained in the union of the Gerschgorin circles $\mathcal{C}_i(t)$ defined by $|z - a_{ii}| \le t\, r_i$. The circles $\mathcal{C}_i(t)$ grow continuously with t from individual points a_{ii} when $t = 0$ to the Gerschgorin circles of $\mathbf{A}$ when $t = 1$,

so, if the circles in the isolated union $\mathcal{U}$ are centered at $a_{i_1 i_1}, a_{i_2 i_2}, \ldots, a_{i_k i_k}$, then for every $t \in [0,1]$ the union $\mathcal{U}(t) = \mathcal{C}_{i_1}(t) \cup \mathcal{C}_{i_2}(t) \cup \cdots \cup \mathcal{C}_{i_k}(t)$ is disjoint from the union $\overline{\mathcal{U}(t)}$ of the other $n-k$ Gerschgorin circles of $\mathbf{C}(t)$. Since (as mentioned in Example 7.1.3) each eigenvalue $\lambda_i(t)$ of $\mathbf{C}(t)$ also varies continuously with t, each $\lambda_i(t)$ is on a continuous curve Γ_i having one end at $\lambda_i(0) = a_{ii}$ and the other end at $\lambda_i(1) \in \sigma(\mathbf{A})$. But since $\mathcal{U}(t) \cap \overline{\mathcal{U}(t)} = \phi$ for all $t \in [0,1]$, the curves $\Gamma_{i_1}, \Gamma_{i_2}, \ldots, \Gamma_{i_k}$ are entirely contained in $\mathcal{U}$, and hence the end points $\lambda_{i_1}(1), \lambda_{i_2}(1), \ldots, \lambda_{i_k}(1)$ are in $\mathcal{U}$. Similarly, the other $n-k$ eigenvalues of $\mathbf{A}$ are in the union of the complementary set of circles. ∎

Example 7.1.5

Problem: Estimate the eigenvalues of $\mathbf{A} = \begin{pmatrix} 5 & 1 & 1 \\ 0 & 6 & 1 \\ 1 & 0 & -5 \end{pmatrix}$.

- A crude estimate is derived from the bound given in Example 7.1.4 on p. 497. Using the ∞-norm, (7.1.12) says that $|\lambda| \leq \|\mathbf{A}\|_\infty = 7$ for all $\lambda \in \sigma(\mathbf{A})$.
- Better estimates are produced by the Gerschgorin circles in Figure 7.1.2 that are derived from row sums. Statements (7.1.13) and (7.1.14) guarantee that one eigenvalue is in (or on) the circle centered at -5, while the remaining two eigenvalues are in (or on) the larger circle centered at $+5$.

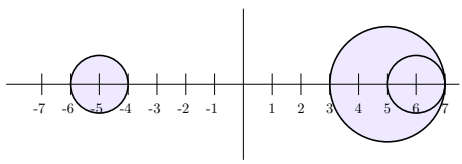


FIGURE 7.1.2. GERSCHGORIN CIRCLES DERIVED FROM ROW SUMS.

- The best estimate is obtained from (7.1.16) by considering $\mathcal{G}_r \cap \mathcal{G}_c$.

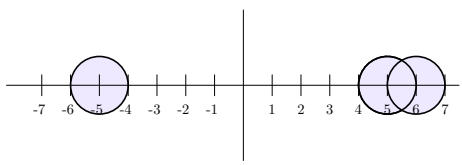


FIGURE 7.1.3. GERSCHGORIN CIRCLES DERIVED FROM $\mathcal{G}_r \cap \mathcal{G}_c$.

In other words, one eigenvalue is in the circle centered at -5, while the other two eigenvalues are in the union of the other two circles in Figure 7.1.3. This is corroborated by computing $\sigma(\mathbf{A}) = \{5,\ (1 \pm 5\sqrt{5})/2\} \approx \{5,\ 6.0902,\ -5.0902\}$.

Example 7.1.6

Diagonally Dominant Matrices Revisited. Recall from Example 4.3.3 on p. 184 that $\mathbf{A}_{n\times n}$ is said to be ***diagonally dominant*** (some authors say *strictly* diagonally dominant) whenever

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \text{for each } i = 1, 2, \ldots, n.$$

Gerschgorin's theorem (7.1.13) guarantees that diagonally dominant matrices cannot possess a zero eigenvalue. But $0 \notin \sigma(\mathbf{A})$ if and only if $\mathbf{A}$ is nonsingular (Exercise 7.1.6), so Gerschgorin's theorem provides an alternative to the argument used in Example 4.3.3 to prove that *all diagonally dominant matrices are nonsingular*.[68] For example, the 3×3 matrix $\mathbf{A}$ in Example 7.1.5 is diagonally dominant, and thus $\mathbf{A}$ is nonsingular. Even when a matrix is not diagonally dominant, Gerschgorin estimates still may be useful in determining whether or not the matrix is nonsingular simply by observing if zero is excluded from $\sigma(\mathbf{A})$ based on the configuration of the Gerschgorin circles given in (7.1.16).

Exercises for section 7.1

7.1.1. Determine the eigenvalues and eigenvectors for the following matrices.

$$\mathbf{A} = \begin{pmatrix} -10 & -7 \\ 14 & 11 \end{pmatrix}. \quad \mathbf{B} = \begin{pmatrix} 2 & 16 & 8 \\ 4 & 14 & 8 \\ -8 & -32 & -18 \end{pmatrix}. \quad \mathbf{C} = \begin{pmatrix} 3 & -2 & 5 \\ 0 & 1 & 4 \\ 0 & -1 & 5 \end{pmatrix}.$$

$$\mathbf{D} = \begin{pmatrix} 0 & 6 & 3 \\ -1 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix}. \quad \mathbf{E} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Which, if any, are deficient in eigenvectors in the sense that there fails to exist a complete linearly independent set?

7.1.2. Without doing an eigenvalue–eigenvector computation, determine which of the following are eigenvectors for

$$\mathbf{A} = \begin{pmatrix} -9 & -6 & -2 & -4 \\ -8 & -6 & -3 & -1 \\ 20 & 15 & 8 & 5 \\ 32 & 21 & 7 & 12 \end{pmatrix},$$

and for those which are eigenvectors, identify the associated eigenvalue.

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

[68] In fact, this result was the motivation behind the original development of Gerschgorin's circles.

7.1.3. Explain why the eigenvalues of triangular and diagonal matrices

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

are simply the diagonal entries—the t_{ii}'s and λ_i's.

7.1.4. For $\mathbf{T} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$, prove $\det(\mathbf{T} - \lambda\mathbf{I}) = \det(\mathbf{A} - \lambda\mathbf{I})\det(\mathbf{C} - \lambda\mathbf{I})$ to conclude that $\sigma\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} = \sigma(\mathbf{A}) \cup \sigma(\mathbf{C})$ for square $\mathbf{A}$ and $\mathbf{C}$.

7.1.5. Determine the eigenvectors of $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. In particular, what is the eigenspace associated with λ_i?

7.1.6. Prove that $0 \in \sigma(\mathbf{A})$ if and only if $\mathbf{A}$ is a singular matrix.

7.1.7. Explain why it's apparent that $\mathbf{A}_{n\times n} = \begin{pmatrix} n & 1 & 1 & \cdots & 1 \\ 1 & n & 1 & \cdots & 1 \\ 1 & 1 & n & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n \end{pmatrix}$ doesn't have a zero eigenvalue, and hence why $\mathbf{A}$ is nonsingular.

7.1.8. Explain why the eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ are real and nonnegative for every $\mathbf{A} \in \mathcal{C}^{m\times n}$. **Hint:** Consider $\|\mathbf{A}\mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2$. When are the eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ strictly positive?

7.1.9. (a) If $\mathbf{A}$ is nonsingular, and if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, show that $(\lambda^{-1}, \mathbf{x})$ is an eigenpair for $\mathbf{A}^{-1}$.

(b) For all $\alpha \notin \sigma(\mathbf{A})$, prove that $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ if and only if $\mathbf{x}$ is an eigenvector of $(\mathbf{A} - \alpha\mathbf{I})^{-1}$.

7.1.10. (a) Show that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(\lambda^k, \mathbf{x})$ is an eigenpair for $\mathbf{A}^k$ for each positive integer k.

(b) If $p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$ is any polynomial, then we define $p(\mathbf{A})$ to be the matrix

$$p(\mathbf{A}) = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} + \alpha_2\mathbf{A}^2 + \cdots + \alpha_k\mathbf{A}^k.$$

Show that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(p(\lambda), \mathbf{x})$ is an eigenpair for $p(\mathbf{A})$.

7.1.11. Explain why (7.1.14) in Gerschgorin's theorem on p. 498 implies that $\mathbf{A} = \begin{pmatrix} 1 & 0 & -2 & 0 \\ 0 & 12 & 0 & -4 \\ 1 & 0 & -1 & 0 \\ 0 & 5 & 0 & 0 \end{pmatrix}$ must have at least two real eigenvalues. Corroborate this fact by computing the eigenvalues of $\mathbf{A}$.

7.1.12. If $\mathbf{A}$ is ***nilpotent*** ($\mathbf{A}^k = \mathbf{0}$ for some k), explain why $trace\,(\mathbf{A}) = 0$. **Hint:** What is $\sigma\,(\mathbf{A})$?

7.1.13. If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are eigenvectors of $\mathbf{A}$ associated with the same eigenvalue λ, explain why every nonzero linear combination

$$\mathbf{v} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_n\mathbf{x}_n$$

is also an eigenvector for $\mathbf{A}$ associated with the eigenvalue λ.

7.1.14. Explain why an eigenvector for a square matrix $\mathbf{A}$ cannot be associated with two distinct eigenvalues for $\mathbf{A}$.

7.1.15. Suppose $\sigma\,(\mathbf{A}_{n\times n}) = \sigma\,(\mathbf{B}_{n\times n})$. Does this guarantee that $\mathbf{A}$ and $\mathbf{B}$ have the same characteristic polynomial?

7.1.16. Construct 2×2 examples to prove the following statements.

(a) $\lambda \in \sigma\,(\mathbf{A})$ and $\mu \in \sigma\,(\mathbf{B}) \;\not\Longrightarrow\; \lambda + \mu \in \sigma\,(\mathbf{A} + \mathbf{B})$.

(b) $\lambda \in \sigma\,(\mathbf{A})$ and $\mu \in \sigma\,(\mathbf{B}) \;\not\Longrightarrow\; \lambda\mu \in \sigma\,(\mathbf{AB})$.

7.1.17. Suppose that $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues for $\mathbf{A}_{n\times n}$, and let $(\lambda_k, \mathbf{c})$ be a particular eigenpair.

(a) For $\lambda \notin \sigma\,(\mathbf{A})$, explain why $(\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{c} = \mathbf{c}/(\lambda_k - \lambda)$.

(b) For an arbitrary vector $\mathbf{d}_{n\times 1}$, prove that the eigenvalues of $\mathbf{A} + \mathbf{c}\mathbf{d}^T$ agree with those of $\mathbf{A}$ except that λ_k is replaced by $\lambda_k + \mathbf{d}^T\mathbf{c}$.

(c) How can $\mathbf{d}$ be selected to guarantee that the eigenvalues of $\mathbf{A} + \mathbf{c}\mathbf{d}^T$ and $\mathbf{A}$ agree except that λ_k is replaced by a specified number μ?

7.1.18. Suppose that $\mathbf{A}$ is a square matrix.

(a) Explain why $\mathbf{A}$ and $\mathbf{A}^T$ have the same eigenvalues.

(b) Explain why $\lambda \in \sigma(\mathbf{A}) \Longleftrightarrow \overline{\lambda} \in \sigma(\mathbf{A}^*)$.
Hint: Recall Exercise 6.1.8.

(c) Do these results imply that $\lambda \in \sigma(\mathbf{A}) \Longleftrightarrow \overline{\lambda} \in \sigma(\mathbf{A})$ when $\mathbf{A}$ is a square matrix of *real* numbers?

(d) A nonzero row vector $\mathbf{y}^*$ is called a ***left-hand*** eigenvector for $\mathbf{A}$ whenever there is a scalar $\mu \in \mathcal{C}$ such that $\mathbf{y}^*(\mathbf{A} - \mu\mathbf{I}) = \mathbf{0}$. Explain why μ must be an eigenvalue for $\mathbf{A}$ in the "right-hand" sense of the term when $\mathbf{A}$ is a square matrix of *real* numbers.

7.1.19. Consider matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times m}$.

(a) Explain why $\mathbf{AB}$ and $\mathbf{BA}$ have the same characteristic polynomial if $m = n$. **Hint:** Recall Exercise 6.2.16.

(b) Explain why the characteristic polynomials for $\mathbf{AB}$ and $\mathbf{BA}$ can't be the same when $m \neq n$, and then explain why $\sigma(\mathbf{AB})$ and $\sigma(\mathbf{BA})$ agree, with the possible exception of a zero eigenvalue.

7.1.20. If $\mathbf{AB} = \mathbf{BA}$, prove that $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvector.

Hint: For $\lambda \in \sigma(\mathbf{A})$, let the columns of $\mathbf{X}$ be a basis for $N(\mathbf{A} - \lambda\mathbf{I})$ so that $(\mathbf{A} - \lambda\mathbf{I})\mathbf{BX} = \mathbf{0}$. Explain why there exists a matrix $\mathbf{P}$ such that $\mathbf{BX} = \mathbf{XP}$, and then consider any eigenpair for $\mathbf{P}$.

7.1.21. For fixed matrices $\mathbf{P}_{m\times m}$ and $\mathbf{Q}_{n\times n}$, let $\mathbf{T}$ be the linear operator on $\mathcal{C}^{m\times n}$ defined by $\mathbf{T}(\mathbf{A}) = \mathbf{PAQ}$.

(a) Show that if $\mathbf{x}$ is a right-hand eigenvector for $\mathbf{P}$ and $\mathbf{y}^*$ is a left-hand eigenvector for $\mathbf{Q}$, then $\mathbf{xy}^*$ is an eigenvector for $\mathbf{T}$.

(b) Explain why $trace(\mathbf{T}) = trace(\mathbf{P})\, trace(\mathbf{Q})$.

7.1.22. Let $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ be a diagonal real matrix such that $\lambda_1 < \lambda_2 < \cdots < \lambda_n$, and let $\mathbf{v}_{n\times 1}$ be a column of real nonzero numbers.

(a) Prove that if α is real and nonzero, then λ_i is not an eigenvalue for $\mathbf{D} + \alpha\mathbf{vv}^T$. Show that the eigenvalues of $\mathbf{D} + \alpha\mathbf{vv}^T$ are in fact given by the solutions of the *secular equation* $f(\xi) = 0$ defined by

$$f(\xi) = 1 + \alpha \sum_{i=1}^{n} \frac{v_i^2}{\lambda_i - \xi}.$$

For $n = 4$ and $\alpha > 0$, verify that the graph of $f(\xi)$ is as depicted in Figure 7.1.4, and thereby conclude that the eigenvalues of $\mathbf{D} + \alpha\mathbf{vv}^T$ interlace with those of $\mathbf{D}$.

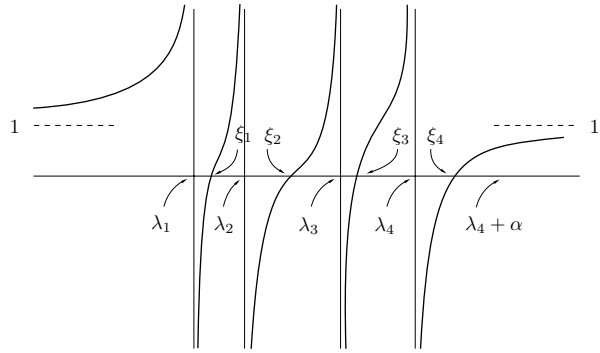


FIGURE 7.1.4

(b) Verify that $(\mathbf{D} - \xi_i \mathbf{I})^{-1}\mathbf{v}$ is an eigenvector for $\mathbf{D} + \alpha \mathbf{v}\mathbf{v}^T$ that is associated with the eigenvalue ξ_i.

7.1.23. Newton's Identities. Let $\lambda_1, \ldots, \lambda_n$ be the roots of the polynomial $p(\lambda) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n$, and let $\tau_k = \lambda_1^k + \lambda_2^k + \cdots + \lambda_n^k$. Newton's identities say $c_k = -(\tau_1 c_{k-1} + \tau_2 c_{k-2} + \cdots + \tau_{k-1}c_1 + \tau_k)/k$. Derive these identities by executing the following steps:

(a) Show $p'(\lambda) = p(\lambda)\sum_{i=1}^n (\lambda - \lambda_i)^{-1}$ (logarithmic differentiation).

(b) Use the geometric series expansion for $(\lambda - \lambda_i)^{-1}$ to show that for $|\lambda| > \max_i |\lambda_i|$,

$$\sum_{i=1}^n \frac{1}{(\lambda - \lambda_i)} = \frac{n}{\lambda} + \frac{\tau_1}{\lambda^2} + \frac{\tau_2}{\lambda^3} + \cdots.$$

(c) Combine these two results, and equate like powers of λ.

7.1.24. Leverrier–Souriau–Frame Algorithm.[69] Let the characteristic equation for $\mathbf{A}$ be given by $\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n = 0$, and define a sequence by taking $\mathbf{B}_0 = \mathbf{I}$ and

$$\mathbf{B}_k = -\frac{trace\,(\mathbf{A}\mathbf{B}_{k-1})}{k}\mathbf{I} + \mathbf{A}\mathbf{B}_{k-1} \quad \text{for} \quad k = 1, 2, \ldots, n.$$

Prove that for each k,

$$c_k = -\frac{trace\,(\mathbf{A}\mathbf{B}_{k-1})}{k}.$$

Hint: Use Newton's identities, and recall Exercise 7.1.10(a).

[69] This algorithm has been rediscovered and modified several times. In 1840, the Frenchman U. J. J. Leverrier provided the basic connection with Newton's identities. J. M. Souriau, also from France, and J. S. Frame, from Michigan State University, independently modified the algorithm to its present form—Souriau's formulation was published in France in 1948, and Frame's method appeared in the United States in 1949. Paul Horst (USA, 1935) along with Faddeev and Sominskii (USSR, 1949) are also credited with rediscovering the technique. Although the algorithm is intriguingly beautiful, it is not practical for floating-point computations.

7.2 DIAGONALIZATION BY SIMILARITY TRANSFORMATIONS

The correct choice of a coordinate system (or basis) often can simplify the form of an equation or the analysis of a particular problem. For example, consider the obliquely oriented ellipse in Figure 7.2.1 whose equation in the xy-coordinate system is

$$13x^2 + 10xy + 13y^2 = 72.$$

By rotating the xy-coordinate system counterclockwise through an angle of $45°$

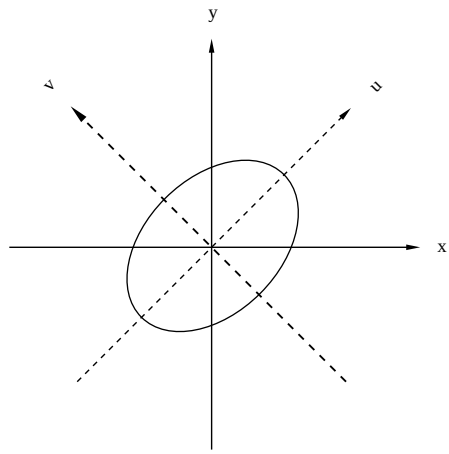


FIGURE 7.2.1

into a uv-coordinate system by means of (5.6.13) on p. 326, the cross-product term is eliminated, and the equation of the ellipse simplifies to become

$$\frac{u^2}{9} + \frac{v^2}{4} = 1.$$

It's shown in Example 7.6.3 on p. 567 that we can do a similar thing for quadratic equations in $\Re^n$.

Choosing or changing to the most appropriate coordinate system (or basis) is always desirable, but in linear algebra it is fundamental. For a linear operator $\mathbf{L}$ on a finite-dimensional space $\mathcal{V}$, the goal is to find a basis $\mathcal{B}$ for $\mathcal{V}$ such that the matrix representation of $\mathbf{L}$ with respect to $\mathcal{B}$ is as simple as possible. Since different matrix representations $\mathbf{A}$ and $\mathbf{B}$ of $\mathbf{L}$ are related by a similarity transformation $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$ (recall §4.8),[70] the fundamental problem for linear operators is strictly a matrix issue—i.e., find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is as simple as possible. The concept of similarity was first introduced on p. 255, but in the interest of continuity it is reviewed below.

[70] While it is helpful to have covered the topics in §§4.7–4.9, much of the subsequent development is accessible without an understanding of this material.

Similarity

- Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be ***similar*** whenever there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$. The product $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is called a ***similarity transformation*** on $\mathbf{A}$.
- **A Fundamental Problem.** Given a square matrix $\mathbf{A}$, reduce it to the simplest possible form by means of a similarity transformation.

Diagonal matrices have the simplest form, so we first ask, "Is every square matrix similar to a diagonal matrix?" Linear algebra and matrix theory would be simpler subjects if this were true, but it's not. For example, consider

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{7.2.1}$$

and observe that $\mathbf{A}^2 = \mathbf{0}$ ($\mathbf{A}$ is nilpotent). If there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$, where $\mathbf{D}$ is diagonal, then

$$\mathbf{D}^2 = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}^2\mathbf{P} = \mathbf{0} \implies \mathbf{D} = \mathbf{0} \implies \mathbf{A} = \mathbf{0},$$

which is false. Thus $\mathbf{A}$, as well as any other nonzero nilpotent matrix, is not similar to a diagonal matrix. Nonzero nilpotent matrices are not the only ones that can't be diagonalized, but, as we will see, nilpotent matrices play a particularly important role in nondiagonalizability.

So, if not all square matrices can be diagonalized by a similarity transformation, what are the characteristics of those that can? An answer is easily derived by examining the equation

$$\mathbf{P}^{-1}\mathbf{A}_{n\times n}\mathbf{P} = \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

which implies $\mathbf{A}\,[\mathbf{P}_{*1} \,|\, \cdots \,|\, \mathbf{P}_{*n}] = [\mathbf{P}_{*1} \,|\, \cdots \,|\, \mathbf{P}_{*n}] \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$ or, equivalently, $[\mathbf{A}\mathbf{P}_{*1} \,|\, \cdots \,|\, \mathbf{A}\mathbf{P}_{*n}] = [\lambda_1\mathbf{P}_{*1} \,|\, \cdots \,|\, \lambda_n\mathbf{P}_{*n}]$. Consequently, $\mathbf{A}\mathbf{P}_{*j} = \lambda_j\mathbf{P}_{*j}$ for each j, so each $(\lambda_j, \mathbf{P}_{*j})$ is an eigenpair for $\mathbf{A}$. In other words, $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$ implies that $\mathbf{P}$ must be a matrix whose columns constitute n linearly independent eigenvectors, and $\mathbf{D}$ is a diagonal matrix whose diagonal entries are the corresponding eigenvalues. It's straightforward to reverse the above argument to prove the converse—i.e., if there exists a linearly independent set of n eigenvectors that are used as columns to build a nonsingular matrix $\mathbf{P}$, and if $\mathbf{D}$ is the diagonal matrix whose diagonal entries are the corresponding eigenvalues, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$. Below is a summary.

Diagonalizability

- A square matrix $\mathbf{A}$ is said to be ***diagonalizable*** whenever $\mathbf{A}$ is similar to a diagonal matrix.
- A ***complete set of eigenvectors*** for $\mathbf{A}_{n\times n}$ is any set of n linearly independent eigenvectors for $\mathbf{A}$. Not all matrices have complete sets of eigenvectors—e.g., consider (7.2.1) or Example 7.1.2. Matrices that fail to possess complete sets of eigenvectors are sometimes called ***deficient*** or ***defective*** matrices.
- $\mathbf{A}_{n\times n}$ is diagonalizable if and only if $\mathbf{A}$ possesses a complete set of eigenvectors. Moreover, $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ if and only if the columns of $\mathbf{P}$ constitute a complete set of eigenvectors and the λ_j 's are the associated eigenvalues—i.e., each $(\lambda_j, \mathbf{P}_{*j})$ is an eigenpair for $\mathbf{A}$.

Example 7.2.1

Problem: If possible, diagonalize the following matrix with a similarity transformation:

$$\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}.$$

Solution: Determine whether or not $\mathbf{A}$ has a complete set of three linearly independent eigenvectors. The characteristic equation—perhaps computed by using (7.1.5)—is

$$\lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2 = 0.$$

Therefore, $\lambda = 1$ is a simple eigenvalue, and $\lambda = -3$ is repeated twice (we say its algebraic multiplicity is 2). Bases for the eigenspaces $N(\mathbf{A} - 1\mathbf{I})$ and $N(\mathbf{A} + 3\mathbf{I})$ are determined in the usual way to be

$$N(\mathbf{A} - 1\mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}\right\} \quad \text{and} \quad N(\mathbf{A} + 3\mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\right\},$$

and it's easy to check that when combined these three eigenvectors constitute a linearly independent set. Consequently, $\mathbf{A}$ must be diagonalizable. To explicitly exhibit the similarity transformation that diagonalizes $\mathbf{A}$, set

$$\mathbf{P} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}, \quad \text{and verify} \quad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & -3 \end{pmatrix} = \mathbf{D}.$$

Since not all square matrices are diagonalizable, it's natural to inquire about the next best thing—i.e., can every square matrix be *triangularized* by similarity? This time the answer is *yes,* but before explaining why, we need to make the following observation.

Similarity Preserves Eigenvalues

Row reductions don't preserve eigenvalues (try a simple example). However, similar matrices have the same characteristic polynomial, so they have the same eigenvalues with the same multiplicities. **Caution!** Similar matrices need not have the same eigenvectors—see Exercise 7.2.3.

Proof. Use the product rule for determinants in conjunction with the fact that $\det\left(\mathbf{P}^{-1}\right) = 1/\det(\mathbf{P})$ (Exercise 6.1.6) to write

$$\begin{aligned}\det(\mathbf{A}-\lambda\mathbf{I}) &= \det\left(\mathbf{P}^{-1}\mathbf{B}\mathbf{P}-\lambda\mathbf{I}\right) = \det\left(\mathbf{P}^{-1}(\mathbf{B}-\lambda\mathbf{I})\mathbf{P}\right)\\ &= \det\left(\mathbf{P}^{-1}\right)\det(\mathbf{B}-\lambda\mathbf{I})\det(\mathbf{P}) = \det(\mathbf{B}-\lambda\mathbf{I}). \quad \blacksquare\end{aligned}$$

In the context of linear operators, this means that the eigenvalues of a matrix representation of an operator $\mathbf{L}$ are invariant under a change of basis. In other words, the eigenvalues are intrinsic to $\mathbf{L}$ in the sense that they are independent of any coordinate representation.

Now we can establish the fact that every square matrix can be triangularized by a similarity transformation. In fact, as Issai Schur (p. 123) realized in 1909, the similarity transformation always can be made to be unitary.

Schur's Triangularization Theorem

Every square matrix is unitarily similar to an upper-triangular matrix. That is, for each $\mathbf{A}_{n\times n}$, there exists a unitary matrix $\mathbf{U}$ (not unique) and an upper-triangular matrix $\mathbf{T}$ (not unique) such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$, and the diagonal entries of $\mathbf{T}$ are the eigenvalues of $\mathbf{A}$.

Proof. Use induction on n, the size of the matrix. For $n = 1$, there is nothing to prove. For $n > 1$, assume that all $n-1 \times n-1$ matrices are unitarily similar to an upper-triangular matrix, and consider an $n \times n$ matrix $\mathbf{A}$. Suppose that $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, and suppose that $\mathbf{x}$ has been normalized so that $\|\mathbf{x}\|_2 = 1$. As discussed on p. 325, we can construct an elementary reflector $\mathbf{R} = \mathbf{R}^* = \mathbf{R}^{-1}$ with the property that $\mathbf{R}\mathbf{x} = \mathbf{e}_1$ or, equivalently, $\mathbf{x} = \mathbf{R}\mathbf{e}_1$ (set $\mathbf{R} = \mathbf{I}$ if $\mathbf{x} = \mathbf{e}_1$). Thus $\mathbf{x}$ is the first column in $\mathbf{R}$, so $\mathbf{R} = \left(\mathbf{x} \,|\, \mathbf{V}\right)$, and

$$\mathbf{R}\mathbf{A}\mathbf{R} = \mathbf{R}\mathbf{A}\left(\mathbf{x} \,|\, \mathbf{V}\right) = \mathbf{R}\left(\lambda\mathbf{x} \,|\, \mathbf{A}\mathbf{V}\right) = \left(\lambda\mathbf{e}_1 \,|\, \mathbf{R}\mathbf{A}\mathbf{V}\right) = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{A}\mathbf{V} \\ \mathbf{0} & \mathbf{V}^*\mathbf{A}\mathbf{V} \end{pmatrix}.$$

Since $\mathbf{V}^*\mathbf{AV}$ is $n-1 \times n-1$, the induction hypothesis insures that there exists a unitary matrix $\mathbf{Q}$ such that $\mathbf{Q}^*(\mathbf{V}^*\mathbf{AV})\mathbf{Q} = \tilde{\mathbf{T}}$ is upper triangular. If $\mathbf{U} = \mathbf{R}\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix}$, then $\mathbf{U}$ is unitary (because $\mathbf{U}^* = \mathbf{U}^{-1}$), and

$$\mathbf{U}^*\mathbf{AU} = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{AVQ} \\ \mathbf{0} & \mathbf{Q}^*\mathbf{V}^*\mathbf{AVQ} \end{pmatrix} = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{AVQ} \\ \mathbf{0} & \tilde{\mathbf{T}} \end{pmatrix} = \mathbf{T}$$

is upper triangular. Since similar matrices have the same eigenvalues, and since the eigenvalues of a triangular matrix are its diagonal entries (Exercise 7.1.3), the diagonal entries of $\mathbf{T}$ must be the eigenvalues of $\mathbf{A}$. ∎

Example 7.2.2

The Cayley–Hamilton[71] **theorem** asserts that every square matrix satisfies its own characteristic equation $p(\lambda) = 0$. That is, $p(\mathbf{A}) = \mathbf{0}$.

Problem: Show how the Cayley–Hamilton theorem follows from Schur's triangularization theorem.

Solution: Schur's theorem insures the existence of a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{AU} = \mathbf{T}$ is triangular, and the development allows for the eigenvalues $\mathbf{A}$ to appear in any given order on the diagonal of $\mathbf{T}$. So, if $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ with λ_i repeated a_i times, then there is a unitary $\mathbf{U}$ such that

$$\mathbf{U}^*\mathbf{AU} = \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \star & \cdots & \star \\ & \mathbf{T}_2 & \cdots & \star \\ & & \ddots & \vdots \\ & & & \mathbf{T}_k \end{pmatrix}, \quad \text{where} \quad \mathbf{T}_i = \begin{pmatrix} \lambda_i & \star & \cdots & \star \\ & \lambda_i & \cdots & \star \\ & & \ddots & \vdots \\ & & & \lambda_i \end{pmatrix}_{a_i \times a_i}.$$

Consequently, $(\mathbf{T}_i - \lambda_i\mathbf{I})^{a_i} = \mathbf{0}$, so $(\mathbf{T} - \lambda_i\mathbf{I})^{a_i}$ has the form

$$(\mathbf{T} - \lambda_i\mathbf{I})^{a_i} = \begin{pmatrix} \star & \cdots & \star & \cdots & \star \\ & \ddots & \vdots & & \vdots \\ & & \mathbf{0} & \cdots & \star \\ & & & \ddots & \vdots \\ & & & & \star \end{pmatrix} \longleftarrow i^{th} \text{ row of blocks.}$$

[71] William Rowan Hamilton (1805–1865), an Irish mathematical astronomer, established this result in 1853 for his quaternions, matrices of the form $\begin{pmatrix} a + b\mathrm{i} & c + d\mathrm{i} \\ -c + d\mathrm{i} & a - b\mathrm{i} \end{pmatrix}$ that resulted from his attempt to generalize complex numbers. In 1858 Arthur Cayley (p. 80) enunciated the general result, but his argument was simply to make direct computations for 2×2 and 3×3 matrices. Cayley apparently didn't appreciate the subtleties of the result because he stated that a formal proof "was not necessary." Hamilton's quaternions took shape in his mind while walking with his wife along the Royal Canal in Dublin, and he was so inspired that he stopped to carve his idea in the stone of the Brougham Bridge. He believed quaternions would revolutionize mathematical physics, and he spent the rest of his life working on them. But the world did not agree. Hamilton became an unhappy man addicted to alcohol who is reported to have died from a severe attack of gout.

This form insures that $(\mathbf{T}-\lambda_1\mathbf{I})^{a_1}(\mathbf{T}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{T}-\lambda_k\mathbf{I})^{a_k}=\mathbf{0}$. The characteristic equation for $\mathbf{A}$ is $p(\lambda)=(\lambda-\lambda_1)^{a_1}(\lambda-\lambda_2)^{a_2}\cdots(\lambda-\lambda_k)^{a_k}=0$, so

$$\begin{aligned}\mathbf{U}^*p(\mathbf{A})\mathbf{U} &= \mathbf{U}^*(\mathbf{A}-\lambda_1\mathbf{I})^{a_1}(\mathbf{A}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{A}-\lambda_k\mathbf{I})^{a_k}\mathbf{U}\\ &= (\mathbf{T}-\lambda_1\mathbf{I})^{a_1}(\mathbf{T}-\lambda_2\mathbf{I})^{a_2}\cdots(\mathbf{T}-\lambda_k\mathbf{I})^{a_k}=\mathbf{0},\end{aligned}$$

and thus $p(\mathbf{A})=\mathbf{0}$. **Note:** A completely different approach to the Cayley–Hamilton theorem is discussed on p. 532.

Schur's theorem is not the complete story on triangularizing by similarity. By allowing nonunitary similarity transformations, the structure of the upper-triangular matrix $\mathbf{T}$ can be simplified to contain zeros everywhere except on the diagonal and the superdiagonal (the diagonal immediately above the main diagonal). This is the Jordan form developed on p. 590, but some of the seeds are sown here.

Multiplicities

For $\lambda\in\sigma(\mathbf{A})=\{\lambda_1,\lambda_2,\ldots,\lambda_s\}$, we adopt the following definitions.

- The ***algebraic multiplicity*** of λ is the number of times it is repeated as a root of the characteristic polynomial. In other words, $alg\ mult_{\mathbf{A}}(\lambda_i)=a_i$ if and only if $(x-\lambda_1)^{a_1}\cdots(x-\lambda_s)^{a_s}=0$ is the characteristic equation for $\mathbf{A}$.
- When $alg\ mult_{\mathbf{A}}(\lambda)=1$, λ is called a ***simple eigenvalue.***
- The ***geometric multiplicity*** of λ is $\dim N(\mathbf{A}-\lambda\mathbf{I})$. In other words, $geo\ mult_{\mathbf{A}}(\lambda)$ is the maximal number of linearly independent eigenvectors associated with λ.
- Eigenvalues such that $alg\ mult_{\mathbf{A}}(\lambda)=geo\ mult_{\mathbf{A}}(\lambda)$ are called ***semisimple eigenvalues*** of $\mathbf{A}$. It follows from (7.2.2) on p. 511 that a simple eigenvalue is always semisimple, but not conversely.

Example 7.2.3

The algebraic and geometric multiplicity need not agree. For example, the nilpotent matrix $\mathbf{A}=\begin{pmatrix}0&1\\0&0\end{pmatrix}$ in (7.2.1) has only one distinct eigenvalue, $\lambda=0$, that is repeated twice, so $alg\ mult_{\mathbf{A}}(0)=2$. But

$$\dim N(\mathbf{A}-0\mathbf{I})=\dim N(\mathbf{A})=1 \quad\Longrightarrow\quad geo\ mult_{\mathbf{A}}(0)=1.$$

In other words, there is only one linearly independent eigenvector associated with $\lambda=0$ even though $\lambda=0$ is repeated twice as an eigenvalue.

Example 7.2.3 shows that $geo\ mult_{\mathbf{A}}(\lambda)<alg\ mult_{\mathbf{A}}(\lambda)$ is possible. However, the inequality can never go in the reverse direction.

Multiplicity Inequality

For every $\mathbf{A} \in \mathcal{C}^{n\times n}$, and for each $\lambda \in \sigma(\mathbf{A})$,

$$geo\ mult_{\mathbf{A}}(\lambda) \leq alg\ mult_{\mathbf{A}}(\lambda). \tag{7.2.2}$$

Proof. Suppose $alg\ mult_{\mathbf{A}}(\lambda) = k$. Schur's triangularization theorem (p. 508) insures the existence of a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}_{n\times n}\mathbf{U} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}$, where $\mathbf{T}_{11}$ is a $k \times k$ upper-triangular matrix whose diagonal entries are equal to λ, and $\mathbf{T}_{22}$ is an $n-k \times n-k$ upper-triangular matrix with $\lambda \notin \sigma(\mathbf{T}_{22})$. Consequently, $\mathbf{T}_{22} - \lambda\mathbf{I}$ is nonsingular, so

$$\begin{aligned} rank(\mathbf{A} - \lambda\mathbf{I}) &= rank(\mathbf{U}^*(\mathbf{A} - \lambda\mathbf{I})\mathbf{U}) = rank\begin{pmatrix} \mathbf{T}_{11} - \lambda\mathbf{I} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} - \lambda\mathbf{I} \end{pmatrix} \\ &\geq rank(\mathbf{T}_{22} - \lambda\mathbf{I}) = n - k. \end{aligned}$$

The inequality follows from the fact that the rank of a matrix is at least as great as the rank of any submatrix—recall the result on p. 215. Therefore,

$$alg\ mult_{\mathbf{A}}(\lambda) = k \geq n - rank(\mathbf{A} - \lambda\mathbf{I}) = \dim N(\mathbf{A} - \lambda\mathbf{I}) = geo\ mult_{\mathbf{A}}(\lambda). \quad \blacksquare$$

Determining whether or not $\mathbf{A}_{n\times n}$ is diagonalizable is equivalent to determining whether or not $\mathbf{A}$ has a complete linearly independent set of eigenvectors, and this can be done if you are willing and able to compute all of the eigenvalues and eigenvectors for $\mathbf{A}$. But this brute force approach can be a monumental task. Fortunately, there are some theoretical tools to help determine how many linearly independent eigenvectors a given matrix possesses.

Independent Eigenvectors

Let $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be a set of distinct eigenvalues for $\mathbf{A}$.

- If $\{(\lambda_1, \mathbf{x}_1), (\lambda_2, \mathbf{x}_2), \ldots, (\lambda_k, \mathbf{x}_k)\}$ is a set of eigenpairs for $\mathbf{A}$, then $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ is a linearly independent set. (7.2.3)
- If $\mathcal{B}_i$ is a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ is a linearly independent set. (7.2.4)

Proof of (7.2.3). Suppose $\mathcal{S}$ is a dependent set. If the vectors in $\mathcal{S}$ are arranged so that $\mathcal{M} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r\}$ is a maximal linearly independent subset, then

$$\mathbf{x}_{r+1} = \sum_{i=1}^{r} \alpha_i \mathbf{x}_i,$$

and multiplication on the left by $\mathbf{A} - \lambda_{r+1}\mathbf{I}$ produces

$$\mathbf{0} = \sum_{i=1}^{r} \alpha_i \left(\mathbf{A}\mathbf{x}_i - \lambda_{r+1}\mathbf{x}_i\right) = \sum_{i=1}^{r} \alpha_i \left(\lambda_i - \lambda_{r+1}\right) \mathbf{x}_i.$$

Because $\mathcal{M}$ is linearly independent, $\alpha_i \left(\lambda_i - \lambda_{r+1}\right) = 0$ for each i. Consequently, $\alpha_i = 0$ for each i (because the eigenvalues are distinct), and hence $\mathbf{x}_{r+1} = \mathbf{0}$. But this is impossible because eigenvectors are nonzero. Therefore, the supposition that $\mathcal{S}$ is a dependent set must be false. ∎

Proof of (7.2.4). The result of Exercise 5.9.14 guarantees that $\mathcal{B}$ is linearly independent if and only if

$$\mathcal{M}_j = N\left(\mathbf{A} - \lambda_j\mathbf{I}\right) \cap \Big[N\left(\mathbf{A} - \lambda_1\mathbf{I}\right) + N\left(\mathbf{A} - \lambda_2\mathbf{I}\right) + \cdots + N\left(\mathbf{A} - \lambda_{j-1}\mathbf{I}\right)\Big] = \mathbf{0}$$

for each $j = 1, 2, \ldots, k$. Suppose we have $\mathbf{0} \neq \mathbf{x} \in \mathcal{M}_j$ for some j. Then $\mathbf{A}\mathbf{x} = \lambda_j\mathbf{x}$ and $\mathbf{x} = \mathbf{v}_1 + \mathbf{v}_2 + \cdots + \mathbf{v}_{j-1}$ for $\mathbf{v}_i \in N\left(\mathbf{A} - \lambda_i\mathbf{I}\right)$, which implies

$$\sum_{i=1}^{j-1}(\lambda_i - \lambda_j)\mathbf{v}_i = \sum_{i=1}^{j-1} \lambda_i\mathbf{v}_i - \lambda_j \sum_{i=1}^{j-1} \mathbf{v}_i = \mathbf{A}\mathbf{x} - \lambda_j\mathbf{x} = \mathbf{0}.$$

By (7.2.3), the $\mathbf{v}_i$'s are linearly independent, and hence $\lambda_i - \lambda_j = 0$ for each $i = 1, 2, \ldots, j-1$. But this is impossible because the eigenvalues are distinct. Therefore, $\mathcal{M}_j = \mathbf{0}$ for each j, and thus $\mathcal{B}$ is linearly independent. ∎

These results lead to the following characterization of diagonalizability.

Diagonalizability and Multiplicities

A matrix $\mathbf{A}_{n\times n}$ is diagonalizable if and only if

$$\textit{geo mult}_{\mathbf{A}}\left(\lambda\right) = \textit{alg mult}_{\mathbf{A}}\left(\lambda\right) \tag{7.2.5}$$

for each $\lambda \in \sigma(\mathbf{A})$—i.e., if and only if every eigenvalue is semisimple.

Proof. Suppose $geo\ mult_{\mathbf{A}}(\lambda_i) = alg\ mult_{\mathbf{A}}(\lambda_i) = a_i$ for each eigenvalue λ_i. If there are k distinct eigenvalues, and if $\mathcal{B}_i$ is a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ contains $\sum_{i=1}^{k} a_i = n$ vectors. We just proved in (7.2.4) that $\mathcal{B}$ is a linearly independent set, so $\mathcal{B}$ represents a complete set of linearly independent eigenvectors of $\mathbf{A}$, and we know this insures that $\mathbf{A}$ must be diagonalizable. Conversely, if $\mathbf{A}$ is diagonalizable, and if λ is an eigenvalue for $\mathbf{A}$ with $alg\ mult_{\mathbf{A}}(\lambda) = a$, then there is a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D} = \begin{pmatrix} \lambda\mathbf{I}_{a\times a} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

where $\lambda \notin \sigma(\mathbf{B})$. Consequently,

$$rank\,(\mathbf{A} - \lambda\mathbf{I}) = rank\ \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} - \lambda\mathbf{I} \end{pmatrix}\mathbf{P}^{-1} = rank\,(\mathbf{B} - \lambda\mathbf{I}) = n - a,$$

and thus

$$geo\ mult_{\mathbf{A}}(\lambda) = \dim N(\mathbf{A} - \lambda\mathbf{I}) = n - rank\,(\mathbf{A} - \lambda\mathbf{I}) = a = alg\ mult_{\mathbf{A}}(\lambda). \quad \blacksquare$$

Example 7.2.4

Problem: Determine if either of the following matrices is diagonalizable:

$$\mathbf{A} = \begin{pmatrix} -1 & -1 & -2 \\ 8 & -11 & -8 \\ -10 & 11 & 7 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}.$$

Solution: Each matrix has exactly the same characteristic equation

$$\lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2 = 0,$$

so $\sigma(\mathbf{A}) = \{1, -3\} = \sigma(\mathbf{B})$, where $\lambda = 1$ has algebraic multiplicity 1 and $\lambda = -3$ has algebraic multiplicity 2. Since

$$geo\ mult_{\mathbf{A}}(-3) = \dim N(\mathbf{A} + 3\mathbf{I}) = 1 < alg\ mult_{\mathbf{A}}(-3),$$

$\mathbf{A}$ is *not* diagonalizable. On the other hand,

$$geo\ mult_{\mathbf{B}}(-3) = \dim N(\mathbf{B} + 3\mathbf{I}) = 2 = alg\ mult_{\mathbf{B}}(-3),$$

and $geo\ mult_{\mathbf{B}}(1) = 1 = alg\ mult_{\mathbf{B}}(1)$, so $\mathbf{B}$ *is* diagonalizable.

If $\mathbf{A}_{n\times n}$ happens to have n distinct eigenvalues, then each eigenvalue is simple. This means that $geo\ mult_{\mathbf{A}}(\lambda) = alg\ mult_{\mathbf{A}}(\lambda) = 1$ for each λ, so (7.2.5) produces the following corollary guaranteeing diagonalizability.

Distinct Eigenvalues

If no eigenvalue of $\mathbf{A}$ is repeated, then $\mathbf{A}$ is diagonalizable. (7.2.6)
Caution! The converse is not true—see Example 7.2.4.

Example 7.2.5

Toeplitz[72] **matrices** have constant entries on each diagonal parallel to the main diagonal. For example, a 4×4 Toeplitz matrix $\mathbf{T}$ along with a *tridiagonal* Toeplitz matrix $\mathbf{A}$ are shown below:

$$\mathbf{T} = \begin{pmatrix} t_0 & t_1 & t_2 & t_3 \\ t_{-1} & t_0 & t_1 & t_2 \\ t_{-2} & t_{-1} & t_0 & t_1 \\ t_{-3} & t_{-2} & t_{-1} & t_0 \end{pmatrix}, \qquad \mathbf{A} = \begin{pmatrix} t_0 & t_1 & 0 & 0 \\ t_{-1} & t_0 & t_1 & 0 \\ 0 & t_{-1} & t_0 & t_1 \\ 0 & 0 & t_{-1} & t_0 \end{pmatrix}.$$

Toeplitz structures occur naturally in a variety of applications, and tridiagonal Toeplitz matrices are commonly the result of discretizing differential equation problems—e.g., see §1.4 (p. 18) and Example 7.6.1 (p. 559). The Toeplitz structure is rich in special properties, but tridiagonal Toeplitz matrices are particularly nice because they are among the few nontrivial structures that admit formulas for their eigenvalues and eigenvectors.

Problem: Show that the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{pmatrix} b & a & & & \\ c & b & a & & \\ & \ddots & \ddots & \ddots & \\ & & c & b & a \\ & & & c & b \end{pmatrix}_{n \times n} \qquad \text{with} \qquad a \neq 0 \neq c$$

are given by

$$\lambda_j = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right) \qquad \text{and} \qquad \mathbf{x}_j = \begin{pmatrix} (c/a)^{1/2}\sin\left(1j\pi/(n+1)\right) \\ (c/a)^{2/2}\sin\left(2j\pi/(n+1)\right) \\ (c/a)^{3/2}\sin\left(3j\pi/(n+1)\right) \\ \vdots \\ (c/a)^{n/2}\sin\left(nj\pi/(n+1)\right) \end{pmatrix}$$

[72] Otto Toeplitz (1881–1940) was a professor in Bonn, Germany, but because of his Jewish background he was dismissed from his chair by the Nazis in 1933. In addition to the matrix that bears his name, Toeplitz is known for his general theory of infinite-dimensional spaces developed in the 1930s.

for $j = 1, 2, \ldots, n$, and conclude that $\mathbf{A}$ is diagonalizable.

Solution: For an eigenpair $(\lambda, \mathbf{x})$, the components in $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$ are $cx_{k-1} + (b-\lambda)x_k + ax_{k+1} = 0$, $k = 1, \ldots, n$ with $x_0 = x_{n+1} = 0$ or, equivalently,

$$x_{k+2} + \left(\frac{b-\lambda}{a}\right) x_{k+1} + \left(\frac{c}{a}\right) x_k = 0 \quad \text{for } k = 0, \ldots, n-1 \text{ with } x_0 = x_{n+1} = 0.$$

These are second-order homogeneous difference equations, and solving them is similar to solving analogous differential equations. The technique is to seek solutions of the form $x_k = \xi r^k$ for constants ξ and r. This produces the quadratic equation $r^2 + (b-\lambda)r/a + c/a = 0$ with roots r_1 and r_2, and it can be argued that the general solution of $x_{k+2} + ((b-\lambda)/a)x_{k+1} + (c/a)x_k = 0$ is

$$x_k = \begin{cases} \alpha r_1^k + \beta r_2^k & \text{if } r_1 \neq r_2, \\ \alpha\rho^k + \beta k\rho^k & \text{if } r_1 = r_2 = \rho, \end{cases} \qquad \text{where } \alpha \text{ and } \beta \text{ are arbitrary constants.}$$

For the eigenvalue problem at hand, r_1 and r_2 must be distinct—otherwise $x_k = \alpha\rho^k + \beta k\rho^k$, and $x_0 = x_{n+1} = 0$ implies each $x_k = 0$, which is impossible because $\mathbf{x}$ is an eigenvector. Hence $x_k = \alpha r_1^k + \beta r_2^k$, and $x_0 = x_{n+1} = 0$ yields

$$\left\{\begin{array}{l} 0 = \alpha + \beta \\ 0 = \alpha r_1^{n+1} + \beta r_2^{n+1} \end{array}\right\} \implies \left(\frac{r_1}{r_2}\right)^{n+1} = \frac{-\beta}{\alpha} = 1 \implies \frac{r_1}{r_2} = \mathrm{e}^{\mathrm{i}2\pi j/(n+1)},$$

so $r_1 = r_2\mathrm{e}^{\mathrm{i}2\pi j/(n+1)}$ for some $1 \leq j \leq n$. Couple this with

$$r^2 + \frac{(b-\lambda)r}{a} + \frac{c}{a} = (r - r_1)(r - r_2) \implies \left\{\begin{array}{l} r_1 r_2 = c/a \\ r_1 + r_2 = -(b-\lambda)/a \end{array}\right.$$

to conclude that $r_1 = \sqrt{c/a}\,\mathrm{e}^{\mathrm{i}\pi j/(n+1)}$, $r_2 = \sqrt{c/a}\,\mathrm{e}^{-\mathrm{i}\pi j/(n+1)}$, and

$$\lambda = b + a\sqrt{c/a}\left(\mathrm{e}^{\mathrm{i}\pi j/(n+1)} + \mathrm{e}^{-\mathrm{i}\pi j/(n+1)}\right) = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right).$$

Therefore, the eigenvalues of $\mathbf{A}$ must be given by

$$\lambda_j = b + 2a\sqrt{c/a}\,\cos\left(\frac{j\pi}{n+1}\right), \quad j = 1, 2, \ldots, n.$$

Since these λ_j's are all distinct ($\cos\theta$ is a strictly decreasing function of θ on $(0, \pi)$, and $a \neq 0 \neq c$), $\mathbf{A}$ must be diagonalizable—recall (7.2.6). Finally, the kth component of any eigenvector associated with λ_j satisfies $x_k = \alpha r_1^k + \beta r_2^k$ with $\alpha + \beta = 0$, so

$$x_k = \alpha\left(\frac{c}{a}\right)^{k/2}\left(\mathrm{e}^{\mathrm{i}\pi jk/(n+1)} - \mathrm{e}^{-\mathrm{i}\pi jk/(n+1)}\right) = 2\mathrm{i}\alpha\left(\frac{c}{a}\right)^{k/2}\sin\left(\frac{jk\pi}{n+1}\right).$$

Setting $\alpha = 1/2\mathrm{i}$ yields a particular eigenvector associated with λ_j as

$$\mathbf{x}_j = \begin{pmatrix} (c/a)^{1/2} \sin(1j\pi/(n+1)) \\ (c/a)^{2/2} \sin(2j\pi/(n+1)) \\ (c/a)^{3/2} \sin(3j\pi/(n+1)) \\ \vdots \\ (c/a)^{n/2} \sin(nj\pi/(n+1)) \end{pmatrix}.$$

Because the λ_j's are distinct, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a complete linearly independent set—recall (7.2.3)—so $\mathbf{P} = \big(\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\big)$ diagonalizes $\mathbf{A}$.

It's often the case that a right-hand and left-hand eigenvector for some eigenvalue is known. Rather than starting from scratch to find additional eigenpairs, the known information can be used to reduce or "deflate" the problem to a smaller one as described in the following example.

Example 7.2.6

Deflation. Suppose that right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ for an eigenvalue λ of $\mathbf{A} \in \Re^{n\times n}$ are already known, so $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*$. Furthermore, suppose $\mathbf{y}^*\mathbf{x} \neq 0$—such eigenvectors are guaranteed to exist if λ is simple or if $\mathbf{A}$ is diagonalizable (Exercises 7.2.23 and 7.2.22).

Problem: Use $\mathbf{x}$ and $\mathbf{y}^*$ to deflate the size of the remaining eigenvalue problem.

Solution: Scale $\mathbf{x}$ and $\mathbf{y}^*$ so that $\mathbf{y}^*\mathbf{x} = 1$, and construct $\mathbf{X}_{n\times n-1}$ so that its columns are an orthonormal basis for $\mathbf{y}^\perp$. An easy way of doing this is to build a reflector $\mathbf{R} = \big[\tilde{\mathbf{y}} \,|\, \mathbf{X}\big]$ having $\tilde{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_2$ as its first column as described on p. 325. If $\mathbf{P} = \big[\mathbf{x} \,|\, \mathbf{X}\big]$, then straightforward multiplication shows that

$$\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{y}^* \\ \mathbf{X}^*(\mathbf{I} - \mathbf{x}\mathbf{y}^*) \end{pmatrix} \quad \text{and} \quad \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

where $\mathbf{B} = \mathbf{X}^*\mathbf{A}\mathbf{X}$ is $n-1 \times n-1$. The eigenvalues of $\mathbf{B}$ constitute the remaining eigenvalues of $\mathbf{A}$ (Exercise 7.1.4), and thus an $n \times n$ eigenvalue problem is deflated to become one of size $n-1 \times n-1$.

Note: When $\mathbf{A}$ is symmetric, we can take $\mathbf{x} = \mathbf{y}$ to be an eigenvector with $\|\mathbf{x}\|_2 = 1$, so $\mathbf{P} = \mathbf{R} = \mathbf{R}^{-1}$, and $\mathbf{R}\mathbf{A}\mathbf{R} = \begin{pmatrix} \lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$ in which $\mathbf{B} = \mathbf{B}^T$.

An elegant and more geometrical way of expressing diagonalizability is now presented to help simplify subsequent analyses and pave the way for extensions.

Spectral Theorem for Diagonalizable Matrices

A matrix $\mathbf{A}_{n\times n}$ with spectrum $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ is diagonalizable if and only if there exist matrices $\{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k\}$ such that

$$\mathbf{A} = \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k, \tag{7.2.7}$$

where the $\mathbf{G}_i$'s have the following properties.

- $\mathbf{G}_i$ is the projector onto $N(\mathbf{A} - \lambda_i\mathbf{I})$ along $R(\mathbf{A} - \lambda_i\mathbf{I})$. (7.2.8)
- $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ whenever $i \neq j$. (7.2.9)
- $\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k = \mathbf{I}$. (7.2.10)

The expansion (7.2.7) is known as the ***spectral decomposition*** of $\mathbf{A}$, and the $\mathbf{G}_i$'s are called the ***spectral projectors*** associated with $\mathbf{A}$.

Proof. If $\mathbf{A}$ is diagonalizable, and if $\mathbf{X}_i$ is a matrix whose columns form a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, then $\mathbf{P} = \big(\mathbf{X}_1 \,|\, \mathbf{X}_2 \,|\, \cdots \,|\, \mathbf{X}_k\big)$ is nonsingular. If $\mathbf{P}^{-1}$ is partitioned in a conformable manner, then we must have

$$\begin{aligned}\mathbf{A} = \mathbf{PDP}^{-1} &= \Big(\mathbf{X}_1 \,|\, \mathbf{X}_2 \,|\, \cdots \,|\, \mathbf{X}_k\Big)\begin{pmatrix}\lambda_1\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \lambda_2\mathbf{I} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \lambda_k\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{Y}_1^T\\ \hline \mathbf{Y}_2^T\\ \hline \vdots\\ \hline \mathbf{Y}_k^T\end{pmatrix} \qquad (7.2.11)\\ &= \lambda_1\mathbf{X}_1\mathbf{Y}_1^T + \lambda_2\mathbf{X}_2\mathbf{Y}_2^T + \cdots + \lambda_k\mathbf{X}_k\mathbf{Y}_k^T\\ &= \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k.\end{aligned}$$

For $\mathbf{G}_i = \mathbf{X}_i\mathbf{Y}_i^T$, the statement $\mathbf{PP}^{-1} = \mathbf{I}$ translates to $\sum_{i=1}^k \mathbf{G}_i = \mathbf{I}$, and

$$\mathbf{P}^{-1}\mathbf{P} = \mathbf{I} \implies \mathbf{Y}_i^T\mathbf{X}_j = \begin{cases}\mathbf{I} & \text{when } i = j,\\ \mathbf{0} & \text{when } i \neq j,\end{cases} \implies \begin{cases}\mathbf{G}_i^2 = \mathbf{G}_i,\\ \mathbf{G}_i\mathbf{G}_j = \mathbf{0} & \text{when } i \neq j.\end{cases}$$

To establish that $R(\mathbf{G}_i) = N(\mathbf{A} - \lambda_i\mathbf{I})$, use $R(\mathbf{AB}) \subseteq R(\mathbf{A})$ (Exercise 4.2.12) and $\mathbf{Y}_i^T\mathbf{X}_i = \mathbf{I}$ to write

$$R(\mathbf{G}_i) = R(\mathbf{X}_i\mathbf{Y}_i^T) \subseteq R(\mathbf{X}_i) = R(\mathbf{X}_i\mathbf{Y}_i^T\mathbf{X}_i) = R(\mathbf{G}_i\mathbf{X}_i) \subseteq R(\mathbf{G}_i).$$

Thus $R(\mathbf{G}_i) = R(\mathbf{X}_i) = N(\mathbf{A} - \lambda_i\mathbf{I})$. To show $N(\mathbf{G}_i) = R(\mathbf{A} - \lambda_i\mathbf{I})$, use $\mathbf{A} = \sum_{j=1}^k \lambda_j\mathbf{G}_j$ with the already established properties of the $\mathbf{G}_i$'s to conclude

$$\mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I}) = \mathbf{G}_i\left(\sum_{j=1}^k \lambda_j\mathbf{G}_j - \lambda_i\sum_{j=1}^k \mathbf{G}_j\right) = \mathbf{0} \implies R(\mathbf{A} - \lambda_i\mathbf{I}) \subseteq N(\mathbf{G}_i).$$

But we already know that $N(\mathbf{A}-\lambda_i\mathbf{I}) = R(\mathbf{G}_i)$, so

$$\dim R(\mathbf{A}-\lambda_i\mathbf{I}) = n - \dim N(\mathbf{A}-\lambda_i\mathbf{I}) = n - \dim R(\mathbf{G}_i) = \dim N(\mathbf{G}_i),$$

and therefore, by (4.4.6), $R(\mathbf{A}-\lambda_i\mathbf{I}) = N(\mathbf{G}_i)$. Conversely, if there exist matrices $\mathbf{G}_i$ satisfying (7.2.8)–(7.2.10), then $\mathbf{A}$ must be diagonalizable. To see this, note that (7.2.8) insures $\dim R(\mathbf{G}_i) = \dim N(\mathbf{A}-\lambda_i\mathbf{I}) = \textit{geo mult}_{\mathbf{A}}(\lambda_i)$, while (7.2.9) implies $R(\mathbf{G}_i) \cap R(\mathbf{G}_j) = \mathbf{0}$ and $R\big(\sum_{i=1}^k \mathbf{G}_i\big) = \sum_{i=1}^k R(\mathbf{G}_i)$ (Exercise 5.9.17). Use these with (7.2.10) in the formula for the dimension of a sum (4.4.19) to write

$$\begin{aligned} n = \dim R(\mathbf{I}) &= \dim R(\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k) \\ &= \dim [R(\mathbf{G}_1) + R(\mathbf{G}_2) + \cdots + R(\mathbf{G}_k)] \\ &= \dim R(\mathbf{G}_1) + \dim R(\mathbf{G}_2) + \cdots + \dim R(\mathbf{G}_k) \\ &= \textit{geo mult}_{\mathbf{A}}(\lambda_1) + \textit{geo mult}_{\mathbf{A}}(\lambda_2) + \cdots + \textit{geo mult}_{\mathbf{A}}(\lambda_k). \end{aligned}$$

Since $\textit{geo mult}_{\mathbf{A}}(\lambda_i) \le \textit{alg mult}_{\mathbf{A}}(\lambda_i)$ and $\sum_{i=1}^k \textit{alg mult}_{\mathbf{A}}(\lambda_i) = n$, the above equation insures that $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i)$ for each i, and, by (7.2.5), this means $\mathbf{A}$ is diagonalizable. ∎

Simple Eigenvalues and Projectors

If $\mathbf{x}$ and $\mathbf{y}^*$ are respective right-hand and left-hand eigenvectors associated with a *simple* eigenvalue $\lambda \in \sigma(\mathbf{A})$, then

$$\mathbf{G} = \mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x} \tag{7.2.12}$$

is the projector onto $N(\mathbf{A}-\lambda\mathbf{I})$ along $R(\mathbf{A}-\lambda\mathbf{I})$. In the context of the spectral theorem (p. 517), this means that $\mathbf{G}$ is the spectral projector associated with λ.

Proof. It's not difficult to prove $\mathbf{y}^*\mathbf{x} \neq 0$ (Exercise 7.2.23), and it's clear that $\mathbf{G}$ is a projector because $\mathbf{G}^2 = \mathbf{x}(\mathbf{y}^*\mathbf{x})\mathbf{y}^*/(\mathbf{y}^*\mathbf{x})^2 = \mathbf{G}$. Now determine $R(\mathbf{G})$. The image of any $\mathbf{z}$ is $\mathbf{G}\mathbf{z} = \alpha\mathbf{x}$ with $\alpha = \mathbf{y}^*\mathbf{z}/\mathbf{y}^*\mathbf{x}$, so

$$R(\mathbf{G}) \subseteq \textit{span}\{\mathbf{x}\} = N(\mathbf{A}-\lambda\mathbf{I}) \quad \text{and} \quad \dim R(\mathbf{G}) = 1 = \dim N(\mathbf{A}-\lambda\mathbf{I}).$$

Thus $R(\mathbf{G}) = N(\mathbf{A}-\lambda\mathbf{I})$. To find $N(\mathbf{G})$, recall $N(\mathbf{G}) = R(\mathbf{I}-\mathbf{G})$ (see (5.9.11), p. 386), and observe that $\mathbf{y}^*(\mathbf{A}-\lambda\mathbf{I}) = \mathbf{0} \implies \mathbf{y}^*(\mathbf{I}-\mathbf{G}) = \mathbf{0}$, so

$$R(\mathbf{A}-\lambda\mathbf{I})^\perp \subseteq R(\mathbf{I}-\mathbf{G})^\perp = N(\mathbf{G})^\perp \Longrightarrow N(\mathbf{G}) \subseteq R(\mathbf{A}-\lambda\mathbf{I}) \text{ (Exercise 5.11.5).}$$

But $\dim N(\mathbf{G}) = n - \dim R(\mathbf{G}) = n-1 = n - \dim N(\mathbf{A}-\lambda\mathbf{I}) = \dim R(\mathbf{A}-\lambda\mathbf{I})$, so $N(\mathbf{G}) = R(\mathbf{A}-\lambda\mathbf{I})$. ∎

Example 7.2.7

Problem: Determine the spectral projectors for $\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}$.

Solution: This is the diagonalizable matrix from Example 7.2.1 (p. 507). Since there are two distinct eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = -3$, there are two spectral projectors,

$$\begin{aligned} \mathbf{G}_1 &= \text{ the projector onto } N(\mathbf{A} - 1\mathbf{I}) \text{ along } R(\mathbf{A} - 1\mathbf{I}), \\ \mathbf{G}_2 &= \text{ the projector onto } N(\mathbf{A} + 3\mathbf{I}) \text{ along } R(\mathbf{A} + 3\mathbf{I}). \end{aligned}$$

There are several different ways to find these projectors.

1. Compute bases for the necessary nullspaces and ranges, and use (5.9.12).

2. Compute $\mathbf{G}_i = \mathbf{X}_i\mathbf{Y}_i^T$ as described in (7.2.11). The required computations are essentially the same as those needed above. Since much of the work has already been done in Example 7.2.1, let's complete the arithmetic. We have

$$\mathbf{P} = \left(\begin{array}{c|cc} 1 & 1 & 1 \\ 2 & 1 & 0 \\ -2 & 0 & 1 \end{array}\right) = \Big(\mathbf{X}_1 \,|\, \mathbf{X}_2\Big), \quad \mathbf{P}^{-1} = \left(\begin{array}{ccc} 1 & -1 & -1 \\ \hline -2 & 3 & 2 \\ 2 & -2 & -1 \end{array}\right) = \left(\begin{array}{c} \mathbf{Y}_1^T \\ \hline \mathbf{Y}_2^T \end{array}\right),$$

so

$$\mathbf{G}_1 = \mathbf{X}_1\mathbf{Y}_1^T = \begin{pmatrix} 1 & -1 & -1 \\ 2 & -2 & -2 \\ -2 & 2 & 2 \end{pmatrix}, \quad \mathbf{G}_2 = \mathbf{X}_2\mathbf{Y}_2^T = \begin{pmatrix} 0 & 1 & 1 \\ -2 & 3 & 2 \\ 2 & -2 & -1 \end{pmatrix}.$$

Check that these are correct by confirming the validity of (7.2.7)–(7.2.10).

3. Since $\lambda_1 = 1$ is a simple eigenvalue, (7.2.12) may be used to compute $\mathbf{G}_1$ from any pair of associated right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^T$. Of course, $\mathbf{P}$ and $\mathbf{P}^{-1}$ are not needed to determine such a pair, but since $\mathbf{P}$ and $\mathbf{P}^{-1}$ have been computed above, we can use $\mathbf{X}_1$ and $\mathbf{Y}_1^T$ to make the point that *any* right-hand and left-hand eigenvectors associated with $\lambda_1 = 1$ will do the job because they are all of the form $\mathbf{x} = \alpha\mathbf{X}_1$ and $\mathbf{y}^T = \beta\mathbf{Y}_1^T$ for $\alpha \neq 0 \neq \beta$. Consequently,

$$\mathbf{G}_1 = \frac{\mathbf{x}\mathbf{y}^T}{\mathbf{y}^T\mathbf{x}} = \frac{\alpha \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \beta \begin{pmatrix} 1 & -1 & -1 \end{pmatrix}}{\alpha\beta} = \begin{pmatrix} 1 & -1 & -1 \\ 2 & -2 & -2 \\ -2 & 2 & 2 \end{pmatrix}.$$

Invoking (7.2.10) yields the other spectral projector as $\mathbf{G}_2 = \mathbf{I} - \mathbf{G}_1$.

4. An even easier solution is obtained from the spectral theorem by writing

$$\begin{aligned} \mathbf{A} - \mathbf{I} &= (1\mathbf{G}_1 - 3\mathbf{G}_2) - (\mathbf{G}_1 + \mathbf{G}_2) = -4\mathbf{G}_2, \\ \mathbf{A} + 3\mathbf{I} &= (1\mathbf{G}_1 - 3\mathbf{G}_2) + 3(\mathbf{G}_1 + \mathbf{G}_2) = 4\mathbf{G}_1, \end{aligned}$$

so that

$$\mathbf{G}_1 = \frac{(\mathbf{A} + 3\mathbf{I})}{4} \quad \text{and} \quad \mathbf{G}_2 = \frac{-(\mathbf{A} - \mathbf{I})}{4}.$$

Can you see how to make this rather ad hoc technique work in more general situations?

5. In fact, the technique above is really a special case of a completely general formula giving each $\mathbf{G}_i$ as a function $\mathbf{A}$ and λ_i as

$$\mathbf{G}_i = \frac{\prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I})}{\prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\lambda_i - \lambda_j)}.$$

This "interpolation formula" is developed on p. 529.

Below is a summary of the facts concerning diagonalizability.

Summary of Diagonalizability

For an $n \times n$ matrix $\mathbf{A}$ with spectrum $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, the following statements are equivalent.

- $\mathbf{A}$ is similar to a diagonal matrix—i.e., $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$.
- $\mathbf{A}$ has a complete linearly independent set of eigenvectors.
- Every λ_i is semisimple—i.e., $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i)$.
- $\mathbf{A} = \lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k$, where
  - ▷ $\mathbf{G}_i$ is the projector onto $N(\mathbf{A} - \lambda_i\mathbf{I})$ along $R(\mathbf{A} - \lambda_i\mathbf{I})$,
  - ▷ $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ whenever $i \neq j$,
  - ▷ $\mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_k = \mathbf{I}$,
  - ▷ $\mathbf{G}_i = \prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\mathbf{A} - \lambda_j\mathbf{I}) \Big/ \prod\limits_{\substack{j=1\\ j\neq i}}^{k} (\lambda_i - \lambda_j)$ (see (7.3.11) on p. 529).
  - ▷ If λ_i is a simple eigenvalue associated with right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$, respectively, then $\mathbf{G}_i = \mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x}$.

Exercises for section 7.2

7.2.1. Diagonalize $\mathbf{A} = \begin{pmatrix} -8 & -6 \\ 12 & 10 \end{pmatrix}$ with a similarity transformation, or else explain why $\mathbf{A}$ can't be diagonalized.

7.2.2. (a) Verify that $alg\ mult_{\mathbf{A}}(\lambda) = geo\ mult_{\mathbf{A}}(\lambda)$ for each eigenvalue of

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

(b) Find a nonsingular $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is a diagonal matrix.

7.2.3. Show that similar matrices need not have the same eigenvectors by giving an example of two matrices that are similar but have different eigenspaces.

7.2.4. $\lambda = 2$ is an eigenvalue for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ 0 & 2 & 0 \\ -2 & -3 & 0 \end{pmatrix}$. Find $alg\ mult_{\mathbf{A}}(\lambda)$ as well as $geo\ mult_{\mathbf{A}}(\lambda)$. Can you conclude anything about the diagonalizability of $\mathbf{A}$ from these results?

7.2.5. If $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, explain why $\mathbf{B}^k = \mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}$.

7.2.6. Compute $\lim_{n\to\infty} \mathbf{A}^n$ for $\mathbf{A} = \begin{pmatrix} 7/5 & 1/5 \\ -1 & 1/2 \end{pmatrix}$.

7.2.7. Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$ be a set of linearly independent eigenvectors for $\mathbf{A}_{n\times n}$ associated with respective eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_t\}$, and let $\mathbf{X}$ be any $n \times (n-t)$ matrix such that $\mathbf{P}_{n\times n} = \big(\mathbf{x}_1 \,|\, \cdots \,|\, \mathbf{x}_t \,|\, \mathbf{X}\big)$ is nonsingular. Prove that if $\mathbf{P}^{-1} = \begin{pmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_t^* \\ \mathbf{Y}^* \end{pmatrix}$, where the $\mathbf{y}_i^*$'s are rows and $\mathbf{Y}^*$ is $(n-t)\times n$, then $\{\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_t^*\}$ is a set of linearly independent left-hand eigenvectors associated with $\{\lambda_1, \lambda_2, \ldots, \lambda_t\}$, respectively (i.e., $\mathbf{y}_i^*\mathbf{A} = \lambda_i \mathbf{y}_i^*$).

7.2.8. Let $\mathbf{A}$ be a diagonalizable matrix, and let $\rho(\star)$ denote the spectral radius (recall Example 7.1.4 on p. 497). Prove that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$ if and only if $\rho(\mathbf{A}) < 1$. **Note:** It is demonstrated on p. 617 that this result holds for nondiagonalizable matrices as well.

7.2.9. Apply the technique used to prove Schur's triangularization theorem (p. 508) to construct an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is upper triangular for $\mathbf{A} = \begin{pmatrix} 13 & -9 \\ 16 & -11 \end{pmatrix}$.

7.2.10. Verify the Cayley–Hamilton theorem for $\mathbf{A} = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}$.
Hint: This is the matrix from Example 7.2.1 on p. 507.

7.2.11. Since each row sum in the following symmetric matrix $\mathbf{A}$ is 4, it's clear that $\mathbf{x} = (1,1,1,1)^T$ is both a right-hand and left-hand eigenvector associated with $\lambda = 4 \in \sigma(\mathbf{A})$. Use the deflation technique of Example 7.2.6 (p. 516) to determine the remaining eigenvalues of

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}.$$

7.2.12. Explain why $\mathbf{A}\mathbf{G}_i = \mathbf{G}_i\mathbf{A} = \lambda_i \mathbf{G}_i$ for the spectral projector $\mathbf{G}_i$ associated with the eigenvalue λ_i of a diagonalizable matrix $\mathbf{A}$.

7.2.13. Prove that $\mathbf{A} = \mathbf{c}_{n\times 1}\mathbf{d}^T_{1\times n}$ is diagonalizable if and only if $\mathbf{d}^T\mathbf{c} \neq 0$.

7.2.14. Prove that $\mathbf{A} = \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$ is diagonalizable if and only if $\mathbf{W}_{s\times s}$ and $\mathbf{Z}_{t\times t}$ are each diagonalizable.

7.2.15. Prove that if $\mathbf{AB} = \mathbf{BA}$, then $\mathbf{A}$ and $\mathbf{B}$ can be *simultaneously* triangularized by a unitary similarity transformation—i.e., $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}_1$ and $\mathbf{U}^*\mathbf{B}\mathbf{U} = \mathbf{T}_2$ for some unitary matrix $\mathbf{U}$. **Hint:** Recall Exercise 7.1.20 (p. 503) along with the development of Schur's triangularization theorem (p. 508).

7.2.16. For diagonalizable matrices, prove that $\mathbf{AB} = \mathbf{BA}$ if and only if $\mathbf{A}$ and $\mathbf{B}$ can be *simultaneously* diagonalized—i.e., $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}_1$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \mathbf{D}_2$ for some $\mathbf{P}$. **Hint:** If $\mathbf{A}$ and $\mathbf{B}$ commute, then so do $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$.

7.2.17. Explain why the following "proof" of the Cayley–Hamilton theorem is not valid. $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) \implies p(\mathbf{A}) = \det(\mathbf{A} - \mathbf{A}\mathbf{I}) = \det(\mathbf{0}) = 0$.

7.2.18. Show that the eigenvalues of the finite difference matrix (p. 19)

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{n\times n} \text{ are } \lambda_j = 4\sin^2\frac{j\pi}{2(n+1)},\ 1 \le j \le n.$$

7.2.19. Let $\mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}_{n\times n}$.

(a) Show that $\lambda \in \sigma\left(\mathbf{N}+\mathbf{N}^T\right)$ if and only if $\mathrm{i}\lambda \in \sigma\left(\mathbf{N}-\mathbf{N}^T\right)$.
(b) Explain why $\mathbf{N}+\mathbf{N}^T$ is nonsingular if and only if n is even.
(c) Evaluate $\det\left(\mathbf{N}-\mathbf{N}^T\right)/\det\left(\mathbf{N}+\mathbf{N}^T\right)$ when n is even.

7.2.20. A Toeplitz matrix having the form

$$\mathbf{C} = \begin{pmatrix} c_0 & c_{n-1} & c_{n-2} & \cdots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & c_{n-3} & \cdots & c_0 \end{pmatrix}_{n\times n}$$

is called a ***circulant matrix***. If $p(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1}$, and if $\{1, \xi, \xi^2, \ldots, \xi^{n-1}\}$ are the nth roots of unity, then the results of Exercise 5.8.12 (p. 379) insure that

$$\mathbf{F}_n\mathbf{C}\mathbf{F}_n^{-1} = \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix}$$

in which $\mathbf{F}_n$ is the Fourier matrix of order n. Verify these facts for the circulant below by computing its eigenvalues and eigenvectors directly:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

7.2.21. Suppose that $(\lambda, \mathbf{x})$ and $(\mu, \mathbf{y}^*)$ are right-hand and left-hand eigenpairs for $\mathbf{A} \in \Re^{n\times n}$—i.e., $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \mu\mathbf{y}^*$. Explain why $\mathbf{y}^*\mathbf{x} = 0$ whenever $\lambda \neq \mu$.

7.2.22. Consider $\mathbf{A} \in \Re^{n\times n}$.

(a) Show that if $\mathbf{A}$ is diagonalizable, then there are right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ associated with $\lambda \in \sigma(\mathbf{A})$ such that $\mathbf{y}^*\mathbf{x} \neq 0$ so that we can make $\mathbf{y}^*\mathbf{x} = 1$.
(b) Show that not every right-hand and left-hand eigenvector $\mathbf{x}$ and $\mathbf{y}^*$ associated with $\lambda \in \sigma(\mathbf{A})$ must satisfy $\mathbf{y}^*\mathbf{x} \neq 0$.
(c) Show that (a) need not be true when $\mathbf{A}$ is not diagonalizable.

7.2.23. Consider $\mathbf{A} \in \Re^{n\times n}$ with $\lambda \in \sigma(\mathbf{A})$.

(a) Prove that if λ is simple, then $\mathbf{y}^*\mathbf{x} \neq 0$ for every pair of respective right-hand and left-hand eigenvectors $\mathbf{x}$ and $\mathbf{y}^*$ associated with λ regardless of whether or not $\mathbf{A}$ is diagonalizable. **Hint:** Use the core-nilpotent decomposition on p. 397.

(b) Show that $\mathbf{y}^*\mathbf{x} = 0$ is possible when λ is not simple.

7.2.24. For $\mathbf{A} \in \Re^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, show $\mathbf{A}$ is diagonalizable if and only if $\Re^n = N(\mathbf{A} - \lambda_1\mathbf{I}) \oplus N(\mathbf{A} - \lambda_2\mathbf{I}) \oplus \cdots \oplus N(\mathbf{A} - \lambda_k\mathbf{I})$. **Hint:** Recall Exercise 5.9.14.

7.2.25. **The Real Schur Form.** Schur's triangularization theorem (p. 508) insures that every square matrix $\mathbf{A}$ is unitarily similar to an upper-triangular matrix—say, $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$. But even when $\mathbf{A}$ is real, $\mathbf{U}$ and $\mathbf{T}$ may have to be complex if $\mathbf{A}$ has some complex eigenvalues. However, the matrices (and the arithmetic) can be constrained to be real by settling for a block-triangular result with 2×2 or scalar entries on the diagonal. Prove that for each $\mathbf{A} \in \Re^{n\times n}$ there exists an orthogonal matrix $\mathbf{P} \in \Re^{n\times n}$ and real matrices $\mathbf{B}_{ij}$ such that

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1k} \\ \mathbf{0} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_{kk} \end{pmatrix}, \quad \text{where } \mathbf{B}_{jj} \text{ is } 1\times 1 \text{ or } 2\times 2.$$

If $\mathbf{B}_{jj} = [\lambda_j]$ is 1×1, then $\lambda_j \in \sigma(\mathbf{A})$, and if $\mathbf{B}_{jj}$ is 2×2, then $\sigma(\mathbf{B}_{jj}) = \{\lambda_j, \overline{\lambda}_j\} \subseteq \sigma(\mathbf{A})$.

7.2.26. When $\mathbf{A} \in \Re^{n\times n}$ is diagonalizable by a similarity transformation $\mathbf{S}$, then $\mathbf{S}$ may have to be complex if $\mathbf{A}$ has some complex eigenvalues. Analogous to Exercise 7.2.25, we can stay in the realm of real numbers by settling for a block-diagonal result with 1×1 or 2×2 entries on the diagonal. Prove that if $\mathbf{A} \in \Re^{n\times n}$ is diagonalizable with real eigenvalues $\{\rho_1, \ldots, \rho_r\}$ and complex eigenvalues $\{\lambda_1, \overline{\lambda}_1, \lambda_2, \overline{\lambda}_2, \ldots, \lambda_t, \overline{\lambda}_t\}$ with $2t + r = n$, then there exists a nonsingular $\mathbf{P} \in \Re^{n\times n}$ and $\mathbf{B}_j$'s $\in \Re^{2\times 2}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}_t \end{pmatrix}, \quad \text{where} \quad \mathbf{D} = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_r \end{pmatrix},$$

and where $\mathbf{B}_j$ has eigenvalues λ_j and $\overline{\lambda}_j$.

7.2.27. For $\mathbf{A} \in \mathcal{C}^{n\times n}$, prove that $\mathbf{x}^*\mathbf{A}\mathbf{x} = 0$ for all $\mathbf{x} \in \mathcal{C}^{n\times 1} \Rightarrow \mathbf{A} = \mathbf{0}$. Show that $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ for all $\mathbf{x} \in \Re^{n\times 1} \not\Rightarrow \mathbf{A} = \mathbf{0}$, even if $\mathbf{A}$ is real.

7.3 FUNCTIONS OF DIAGONALIZABLE MATRICES

For square matrices $\mathbf{A}$, what should it mean to write $\sin \mathbf{A}$, $\mathrm{e}^{\mathbf{A}}$, $\ln \mathbf{A}$, etc.? A naive approach might be to simply apply the given function to each entry of $\mathbf{A}$ such as

$$\sin \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \stackrel{?}{=} \begin{pmatrix} \sin a_{11} & \sin a_{12} \\ \sin a_{21} & \sin a_{22} \end{pmatrix}. \tag{7.3.1}$$

But doing so results in matrix functions that fail to have the same properties as their scalar counterparts. For example, since $\sin^2 x + \cos^2 x = 1$ for all scalars x, we would like our definitions of $\sin \mathbf{A}$ and $\cos \mathbf{A}$ to result in the analogous matrix identity $\sin^2 \mathbf{A} + \cos^2 \mathbf{A} = \mathbf{I}$ for all square matrices $\mathbf{A}$. The entrywise approach (7.3.1) clearly fails in this regard.

One way to define matrix functions possessing properties consistent with their scalar counterparts is to use infinite series expansions. For example, consider the exponential function

$$\mathrm{e}^z = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} \cdots. \tag{7.3.2}$$

Formally replacing the scalar argument z by a square matrix $\mathbf{A}$ ($z^0 = 1$ is replaced with $\mathbf{A}^0 = \mathbf{I}$) results in the infinite series of matrices

$$\mathrm{e}^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!} \cdots, \tag{7.3.3}$$

called the *matrix exponential.* While this results in a matrix that has properties analogous to its scalar counterpart, it suffers from the fact that convergence must be dealt with, and then there is the problem of describing the entries in the limit. These issues are handled by deriving a closed form expression for (7.3.3).

If $\mathbf{A}$ is diagonalizable, then $\mathbf{A} = \mathbf{PDP}^{-1} = \mathbf{P}\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)\,\mathbf{P}^{-1}$, and $\mathbf{A}^k = \mathbf{PD}^k\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\left(\lambda_1^k, \ldots, \lambda_n^k\right)\mathbf{P}^{-1}$, so

$$\mathrm{e}^{\mathbf{A}} = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} = \sum_{k=0}^{\infty} \frac{\mathbf{PD}^k\mathbf{P}^{-1}}{k!} = \mathbf{P}\left(\sum_{k=0}^{\infty} \frac{\mathbf{D}^k}{k!}\right)\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\left(\mathrm{e}^{\lambda_1}, \ldots, \mathrm{e}^{\lambda_n}\right)\mathbf{P}^{-1}.$$

In other words, we don't have to use the infinite series (7.3.3) to define $\mathrm{e}^{\mathbf{A}}$. Instead, define $\mathrm{e}^{\mathbf{D}} = \mathrm{diag}\,(\mathrm{e}^{\lambda_1}, \mathrm{e}^{\lambda_2}, \ldots, \mathrm{e}^{\lambda_n})$, and set

$$\mathrm{e}^{\mathbf{A}} = \mathbf{P}\mathrm{e}^{\mathbf{D}}\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\,(\mathrm{e}^{\lambda_1}, \mathrm{e}^{\lambda_2}, \ldots, \mathrm{e}^{\lambda_n})\,\mathbf{P}^{-1}.$$

This idea can be generalized to any function $f(z)$ that is defined on the eigenvalues λ_i of a diagonalizable matrix $\mathbf{A} = \mathbf{PDP}^{-1}$ by defining $f(\mathbf{D})$ to be $f(\mathbf{D}) = \mathrm{diag}\,(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n))$ and by setting

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\,\mathrm{diag}\,(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n))\,\mathbf{P}^{-1}. \tag{7.3.4}$$

At first glance this definition seems to have an edge over the infinite series approach because there are no convergence issues to deal with. But convergence worries have been traded for uniqueness worries. Because $\mathbf{P}$ is not unique, it's not apparent that (7.3.4) is well defined. The eigenvector matrix $\mathbf{P}$ you compute for a given $\mathbf{A}$ need not be the same as the eigenvector matrix I compute, so what insures that your $f(\mathbf{A})$ will be the same as mine? The spectral theorem (p. 517) does. Suppose there are k distinct eigenvalues that are grouped according to repetition, and expand (7.3.4) just as (7.2.11) is expanded to produce

$$f(\mathbf{A}) = \mathbf{PDP}^{-1} = \Big(\mathbf{X}_1|\mathbf{X}_2|\cdots|\mathbf{X}_k\Big)\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_k)\mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{Y}_1^T \\ \hline \mathbf{Y}_2^T \\ \hline \vdots \\ \hline \mathbf{Y}_k^T \end{pmatrix}$$

$$= \sum_{i=1}^{k} f(\lambda_i)\mathbf{X}_i\mathbf{Y}_i^T = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i.$$

Since $\mathbf{G}_i$ is the projector onto $N(\mathbf{A}-\lambda_i\mathbf{I})$ along $R(\mathbf{A}-\lambda_i\mathbf{I})$, $\mathbf{G}_i$ is uniquely determined by $\mathbf{A}$. Therefore, (7.3.4) uniquely defines $f(\mathbf{A})$ regardless of the choice of $\mathbf{P}$. We can now make a formal definition.

Functions of Diagonalizable Matrices

Let $\mathbf{A} = \mathbf{PDP}^{-1}$ be a diagonalizable matrix where the eigenvalues in $\mathbf{D} = \operatorname{diag}(\lambda_1\mathbf{I}, \lambda_2\mathbf{I}, \ldots, \lambda_k\mathbf{I})$ are grouped by repetition. For a function $f(z)$ that is defined at each $\lambda_i \in \sigma(\mathbf{A})$, define

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_k)\mathbf{I} \end{pmatrix}\mathbf{P}^{-1} \tag{7.3.5}$$

$$= f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + \cdots + f(\lambda_k)\mathbf{G}_k, \tag{7.3.6}$$

where $\mathbf{G}_i$ is the ith spectral projector as described on pp. 517, 529. The generalization to nondiagonalizable matrices is on p. 603.

The discussion of matrix functions was initiated by considering infinite series, so, to complete the circle, a formal statement connecting infinite series with (7.3.5) and (7.3.6) is needed. By replacing $\mathbf{A}$ by $\mathbf{PDP}^{-1}$ in $\sum_{n=0}^{\infty} c_n(\mathbf{A}-z_0\mathbf{I})^n$ and expanding the result, the following result is established.

Infinite Series

If $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ converges when $|z - z_0| < r$, and if $|\lambda_i - z_0| < r$ for each eigenvalue λ_i of a diagonalizable matrix $\mathbf{A}$, then

$$f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n. \tag{7.3.7}$$

It can be argued that the matrix series on the right-hand side of (7.3.7) converges if and only if $|\lambda_i - z_0| < r$ for each λ_i, regardless of whether or not $\mathbf{A}$ is diagonalizable. So (7.3.7) serves to define $f(\mathbf{A})$ for functions with series expansions regardless of whether or not $\mathbf{A}$ is diagonalizable. More is said in Example 7.9.3 (p. 605).

Example 7.3.1

Neumann Series Revisited. The function $f(z) = (1-z)^{-1}$ has the geometric series expansion $(1-z)^{-1} = \sum_{k=1}^{\infty} z^k$ that converges if and only if $|z| < 1$. This means that the associated matrix function $f(\mathbf{A}) = (\mathbf{I} - \mathbf{A})^{-1}$ is given by

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k \quad \text{if and only if } |\lambda| < 1 \text{ for all } \lambda \in \sigma(\mathbf{A}). \tag{7.3.8}$$

This is the ***Neumann series*** discussed on p. 126, where it was argued that if $\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0}$, then $(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k$. The two approaches are the same because it turns out that $\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0} \Longleftrightarrow |\lambda| < 1$ for all $\lambda \in \sigma(\mathbf{A})$. This is immediate for diagonalizable matrices, but the nondiagonalizable case is a bit more involved—the complete statement is developed on p. 618. Because $\max_i |\lambda_i| \le \|\mathbf{A}\|$ for all matrix norms (Example 7.1.4, p. 497), a corollary of (7.3.8) is that $(\mathbf{I} - \mathbf{A})^{-1}$ exists and

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k \quad \text{when } \|\mathbf{A}\| < 1 \text{ for any matrix norm.} \tag{7.3.9}$$

Caution! $(\mathbf{I} - \mathbf{A})^{-1}$ can exist without the Neumann series expansion being valid because all that's needed for $\mathbf{I} - \mathbf{A}$ to be nonsingular is $1 \notin \sigma(\mathbf{A})$, while convergence of the Neumann series requires each $|\lambda| < 1$.

Example 7.3.2

Eigenvalue Perturbations. It's often important to understand how the eigenvalues of a matrix are affected by perturbations. In general, this is a complicated issue, but for diagonalizable matrices the problem is more tractable.

Problem: Suppose $\mathbf{B} = \mathbf{A} + \mathbf{E}$, where $\mathbf{A}$ is diagonalizable, and let $\beta \in \sigma(\mathbf{B})$. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, explain why

$$\min_{\lambda_i \in \sigma(\mathbf{A})} |\beta - \lambda_i| \le \kappa(\mathbf{P})\, \|\mathbf{E}\|, \quad \text{where} \quad \kappa(\mathbf{P}) = \|\mathbf{P}\|\, \|\mathbf{P}^{-1}\| \tag{7.3.10}$$

for matrix norms satisfying $\|\mathbf{D}\| = \max_i |\lambda_i|$ (e.g., any standard induced norm).

Solution: Assume $\beta \notin \sigma(\mathbf{A})$—(7.3.10) is trivial if $\beta \in \sigma(\mathbf{A})$—and observe that

$$(\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{B}) = (\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{A} - \mathbf{E}) = \mathbf{I} - (\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}$$

implies that $1 \le \|(\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\|$—otherwise $\mathbf{I} - (\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}$ is nonsingular by (7.3.9), which is impossible because $(\beta\mathbf{I} - \mathbf{B})$ (and hence $(\beta\mathbf{I} - \mathbf{A})^{-1}(\beta\mathbf{I} - \mathbf{B})$ is singular). Consequently,

$$\begin{aligned} 1 &\le \|(\beta\mathbf{I} - \mathbf{A})^{-1}\mathbf{E}\| = \|\mathbf{P}(\beta\mathbf{I} - \mathbf{D})^{-1}\mathbf{P}^{-1}\mathbf{E}\| \le \|\mathbf{P}\|\, \|(\beta\mathbf{I} - \mathbf{D})^{-1}\|\, \|\mathbf{P}^{-1}\|\, \|\mathbf{E}\| \\ &= \kappa(\mathbf{P})\, \|\mathbf{E}\| \max_i |\beta - \lambda_i|^{-1} = \kappa(\mathbf{P})\, \|\mathbf{E}\| \frac{1}{\min_i |\beta - \lambda_i|}, \end{aligned}$$

and this produces (7.3.10). Similar to the case of linear systems (Example 5.12.1, p. 414), the expression $\kappa(\mathbf{P})$ is a *condition number* in the sense that if $\kappa(\mathbf{P})$ is relatively small, then the λ_i's are relatively insensitive, but if $\kappa(\mathbf{P})$ is relatively large, we must be suspicious. **Note:** Because it's a corollary of their 1960 results, the bound (7.3.10) is often referred to as the ***Bauer–Fike bound***.

Infinite series representations can always be avoided because *every function of* $\mathbf{A}_{n\times n}$ *can be expressed as a polynomial in* $\mathbf{A}$. In other words, when $f(\mathbf{A})$ exists, there is a polynomial $p(z)$ such that $p(\mathbf{A}) = f(\mathbf{A})$. This is true for all matrices, but the development here is limited to diagonalizable matrices—nondiagonalizable matrices are treated in Exercise 7.3.7. In the diagonalizable case, $f(\mathbf{A})$ exists if and only if $f(\lambda_i)$ exists for each $\lambda_i \in \sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, and, by (7.3.6), $f(\mathbf{A}) = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i$, where $\mathbf{G}_i$ is the ith spectral projector. Any polynomial $p(z)$ agreeing with $f(z)$ on $\sigma(\mathbf{A})$ does the job because if $p(\lambda_i) = f(\lambda_i)$ for each $\lambda_i \in \sigma(\mathbf{A})$, then

$$p(\mathbf{A}) = \sum_{i=1}^{k} p(\lambda_i)\mathbf{G}_i = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i = f(\mathbf{A}).$$

But is there always a polynomial satisfying $p(\lambda_i) = f(\lambda_i)$ for each $\lambda_i \in \sigma(\mathbf{A})$? Sure—that's what the *Lagrange interpolating polynomial* from Example 4.3.5 (p. 186) does. It's given by

$$p(z) = \sum_{i=1}^{k} \left(f(\lambda_i) \frac{\prod_{\substack{j=1 \\ j\neq i}}^{k} (z - \lambda_j)}{\prod_{\substack{j=1 \\ j\neq i}}^{k} (\lambda_i - \lambda_j)} \right), \text{ so } f(\mathbf{A}) = p(\mathbf{A}) = \sum_{i=1}^{k} \left(f(\lambda_i) \frac{\prod_{\substack{j=1 \\ j\neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I})}{\prod_{\substack{j=1 \\ j\neq i}}^{k} (\lambda_i - \lambda_j)} \right).$$

Using the function $g_i(z) = \begin{cases} 1 & \text{if } z = \lambda_i, \\ 0 & \text{if } z \neq \lambda_i, \end{cases}$ with this representation as well as that in (7.3.6) yields $\prod_{\substack{j=1 \\ j\neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I}) \Big/ \prod_{\substack{j=1 \\ j\neq i}}^{k} (\lambda_i - \lambda_j) = g_i(\mathbf{A}) = \mathbf{G}_i$. For example, if $\sigma(\mathbf{A}_{n\times n}) = \{\lambda_1, \lambda_2, \lambda_3\}$, then $f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + f(\lambda_3)\mathbf{G}_3$ with

$$\mathbf{G}_1 = \frac{(\mathbf{A}-\lambda_2\mathbf{I})(\mathbf{A}-\lambda_3\mathbf{I})}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)}, \quad \mathbf{G}_2 = \frac{(\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_3\mathbf{I})}{(\lambda_2-\lambda_1)(\lambda_2-\lambda_3)}, \quad \mathbf{G}_3 = \frac{(\mathbf{A}-\lambda_1\mathbf{I})(\mathbf{A}-\lambda_2\mathbf{I})}{(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}.$$

Below is a summary of these observations.

Spectral Projectors

If $\mathbf{A}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then the spectral projector onto $N(\mathbf{A} - \lambda_i\mathbf{I})$ along $R(\mathbf{A} - \lambda_i\mathbf{I})$ is given by

$$\mathbf{G}_i = \prod_{\substack{j=1 \\ j\neq i}}^{k} (\mathbf{A} - \lambda_j \mathbf{I}) \Big/ \prod_{\substack{j=1 \\ j\neq i}}^{k} (\lambda_i - \lambda_j) \quad \text{for } i = 1, 2, \ldots, k. \tag{7.3.11}$$

Consequently, if $f(z)$ is defined on $\sigma(\mathbf{A})$, then $f(\mathbf{A}) = \sum_{i=1}^{k} f(\lambda_i)\mathbf{G}_i$ is a polynomial in $\mathbf{A}$ of degree at most $k-1$.

Example 7.3.3

Problem: For a scalar t, determine the matrix exponential $e^{\mathbf{A}t}$, where

$$\mathbf{A} = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix} \quad \text{with } \alpha + \beta \neq 0.$$

Solution 1: The characteristic equation for $\mathbf{A}$ is $\lambda^2 + (\alpha + \beta)\lambda = 0$, so the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 0$ and $\lambda_2 = -(\alpha+\beta)$. Note that $\mathbf{A}$ is diagonalizable

because no eigenvalue is repeated—recall (7.2.6). Using the function $f(z) = \mathrm{e}^{zt}$, the spectral representation (7.3.6) says that

$$\mathrm{e}^{\mathbf{A}t} = f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 = \mathrm{e}^{\lambda_1 t}\mathbf{G}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{G}_2.$$

The spectral projectors $\mathbf{G}_1$ and $\mathbf{G}_2$ are determined from (7.3.11) to be

$$\mathbf{G}_1 = \frac{\mathbf{A} - \lambda_2\mathbf{I}}{-\lambda_2} = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} \quad \text{and} \quad \mathbf{G}_2 = \frac{\mathbf{A}}{\lambda_2} = \frac{1}{\alpha+\beta}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix},$$

so

$$\mathrm{e}^{\mathbf{A}t} = \mathbf{G}_1 + \mathrm{e}^{-(\alpha+\beta)t}\mathbf{G}_2 = \frac{1}{\alpha+\beta}\left[\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} + \mathrm{e}^{-(\alpha+\beta)t}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix}\right].$$

Solution 2: Compute eigenpairs $(\lambda_1, \mathbf{x}_1)$ and $(\lambda_2, \mathbf{x}_2)$, construct $\mathbf{P} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2\right]$, and compute

$$\mathrm{e}^{\mathbf{A}t} = \mathbf{P}\begin{pmatrix} f(\lambda_1) & 0 \\ 0 & f(\lambda_2) \end{pmatrix}\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \mathrm{e}^{\lambda_1 t} & 0 \\ 0 & \mathrm{e}^{\lambda_2 t} \end{pmatrix}\mathbf{P}^{-1}.$$

The computational details are called for in Exercise 7.3.2.

Example 7.3.4

Problem: For $\mathbf{T} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$, evaluate $\lim_{k\to\infty} \mathbf{T}^k$.

Solution 1: Compute two eigenpairs, $\lambda_1 = 1$, $\mathbf{x}_1 = (1,\ 1)^T$, and $\lambda_2 = 1/4$, $\mathbf{x}_2 = (-2,\ 1)^T$. If $\mathbf{P} = [\mathbf{x}_1 \,|\, \mathbf{x}_2]$, then $\mathbf{T} = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 1/4 \end{pmatrix}\mathbf{P}^{-1}$, so

$$\mathbf{T}^k = \mathbf{P}\begin{pmatrix} 1^k & 0 \\ 0 & 1/4^k \end{pmatrix}\mathbf{P}^{-1} \to \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{P}^{-1} = \frac{1}{3}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}. \tag{7.3.12}$$

Solution 2: We know from (7.3.6) that $\mathbf{T}^k = 1^k\mathbf{G}_1 + (1/4)^k\mathbf{G}_2 \to \mathbf{G}_1$. Since $\lambda_1 = 1$ is a simple eigenvalue, formula (7.2.12) on p. 518 can be used to compute $\mathbf{G}_1 = \mathbf{x}_1\mathbf{y}_1^T/\mathbf{y}_1^T\mathbf{x}_1$, where $\mathbf{x}_1$ and $\mathbf{y}_1^T$ are any right- and left-hand eigenvectors associated with $\lambda_1 = 1$. A right-hand eigenvector $\mathbf{x}_1$ was computed above. Computing a left-hand eigenvector $\mathbf{y}_1^T = (1,\ 2)$ yields

$$\mathbf{T}^k \to \mathbf{G}_1 = \frac{\mathbf{x}_1\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1} = \frac{1}{3}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}. \tag{7.3.13}$$

Example 7.3.5

Population Migration. Suppose that the population migration between two geographical regions—say, the North and the South—is as follows. Each year, 50% of the population in the North migrates to the South, while only 25% of the population in the South moves to the North. This situation is depicted by drawing a transition diagram as shown below in Figure 7.3.1.

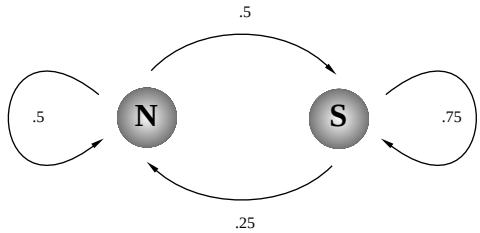


FIGURE 7.3.1

Problem: If this migration pattern continues, will the population in the North continually shrink until the entire population is eventually in the South, or will the population distribution somehow stabilize before the North is completely deserted?

Solution: Let n_k and s_k denote the respective proportions of the total population living in the North and South at the end of year k, and assume $n_k + s_k = 1$. The migration pattern dictates that the fractions of the population in each region at the end of year $k+1$ are

$$\left\{\begin{array}{l} n_{k+1} = n_k(.5) + s_k(.25) \\ s_{k+1} = n_k(.5) + s_k(.75) \end{array}\right\} \quad \text{or, equivalently,} \quad \mathbf{p}_{k+1}^T = \mathbf{p}_k^T\mathbf{T}, \tag{7.3.14}$$

where $\mathbf{p}_k^T = (\,n_k \quad s_k\,)$ and $\mathbf{p}_{k+1}^T = (\,n_{k+1} \quad s_{k+1}\,)$ are the respective population distributions at the end of years k and $k+1$, and where

$$\mathbf{T} = \begin{array}{c} \\ \text{N} \\ \text{S} \end{array}\!\!\begin{array}{c} \begin{array}{cc} \text{N} & \text{S} \end{array} \\ \begin{pmatrix} .5 & .5 \\ .25 & .75 \end{pmatrix} \end{array}$$

is the associated ***transition matrix*** (recall Example 3.6.3). Inducting on

$$\mathbf{p}_1^T = \mathbf{p}_0^T\mathbf{T}, \quad \mathbf{p}_2^T = \mathbf{p}_1^T\mathbf{T} = \mathbf{p}_0^T\mathbf{T}^2, \quad \mathbf{p}_3^T = \mathbf{p}_2^T\mathbf{T} = \mathbf{p}_0^T\mathbf{T}^3, \quad \cdots$$

leads to $\mathbf{p}_k^T = \mathbf{p}_0^T\mathbf{T}^k$, which indicates that the powers of $\mathbf{T}$ determine how the process evolves. Determining the long-run population distribution[73] is therefore

[73] The long-run distribution goes by a lot of different names. It's also called the *limiting* distribution, the *steady-state* distribution, and the *stationary* distribution.

accomplished by analyzing $\lim_{k\to\infty} \mathbf{T}^k$. The results of Example 7.3.4 together with $n_0 + s_0 = 1$ yield the long-run (or limiting) population distribution as

$$\mathbf{p}_\infty^T = \lim_{k\to\infty} \mathbf{p}_k^T = \lim_{k\to\infty} \mathbf{p}_0^T \mathbf{T}^k = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{T}^k = \begin{pmatrix} n_0 & s_0 \end{pmatrix} \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix}$$
$$= \begin{pmatrix} \dfrac{n_0 + s_0}{3} & \dfrac{2(n_0 + s_0)}{3} \end{pmatrix} = \begin{pmatrix} \dfrac{1}{3} & \dfrac{2}{3} \end{pmatrix}.$$

So if the migration pattern continues to hold, then the population distribution will eventually stabilize with 1/3 of the population being in the North and 2/3 of the population in the South. And this is independent of the initial distribution!

Observations: This is an example of a broader class of evolutionary processes known as ***Markov chains*** (p. 687), and the following observations are typical.

- It's clear from (7.3.12) or (7.3.13) that the rate at which the population distribution stabilizes is governed by how fast $(1/4)^k \to 0$. In other words, the magnitude of the largest subdominant eigenvalue of $\mathbf{T}$ determines the rate of evolution.
- For the dominant eigenvalue $\lambda_1 = 1$, the column, $\mathbf{x}_1$, of 1's is a right-hand eigenvector (because $\mathbf{T}$ has unit row sums). This forces the limiting distribution $\mathbf{p}_\infty^T$ to be a particular left-hand eigenvector associated with $\lambda_1 = 1$ because for an arbitrary left-hand eigenvector $\mathbf{y}_1^T$ associated with $\lambda_1 = 1$, equation (7.3.13) in Example 7.3.4 insures that

$$\mathbf{p}_\infty^T = \lim_{k\to\infty} \mathbf{p}_0^T \mathbf{T}^k = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{T}^k = \mathbf{p}_0^T \mathbf{G}_1 = \frac{(\mathbf{p}_0^T \mathbf{x}_1)\mathbf{y}_1^T}{\mathbf{y}_1^T \mathbf{x}_1} = \frac{\mathbf{y}_1^T}{\mathbf{y}_1^T \mathbf{x}_1}. \tag{7.3.15}$$

  The fact that $\mathbf{p}_0^T \mathbf{T}^k$ converges to an eigenvector is a special case of the power method discussed in Example 7.3.7.
- Equation (7.3.15) shows why the initial distribution $\mathbf{p}_0^T$ always drops away in the limit. But $\mathbf{p}_0^T$ is not completely irrelevant because it always affects the transient behavior—i.e., the behavior of $\mathbf{p}_k^T = \mathbf{p}_0^T \mathbf{T}^k$ for smaller k's.

Example 7.3.6

Cayley–Hamilton Revisited. The Cayley–Hamilton theorem (p. 509) says that if $p(\lambda) = 0$ is the characteristic equation for $\mathbf{A}$, then $p(\mathbf{A}) = \mathbf{0}$. This is evident for diagonalizable $\mathbf{A}$ because $p(\lambda_i) = 0$ for each $\lambda_i \in \sigma(\mathbf{A})$, so, by (7.3.6), $p(\mathbf{A}) = p(\lambda_1)\mathbf{G}_1 + p(\lambda_2)\mathbf{G}_2 + \cdots + p(\lambda_k)\mathbf{G}_k = \mathbf{0}$.

Problem: Establish the Cayley–Hamilton theorem for nondiagonalizable matrices by using the diagonalizable result together with a continuity argument.

Solution: Schur's triangularization theorem (p. 508) insures $\mathbf{A}_{n\times n} = \mathbf{UTU}^*$ for a unitary $\mathbf{U}$ and an upper triangular $\mathbf{T}$ having the eigenvalues of $\mathbf{A}$ on the

diagonal. For each $\epsilon \neq 0$, it's possible to find numbers ϵ_i such that $(\lambda_1 + \epsilon_1)$, $(\lambda_2 + \epsilon_2)$, $\ldots$, $(\lambda_n + \epsilon_n)$ are distinct and $\sum \epsilon_i^2 = |\epsilon|$. Set

$$\mathbf{D}(\epsilon) = \operatorname{diag}\,(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) \quad \text{and} \quad \mathbf{B}(\epsilon) = \mathbf{U}\big(\mathbf{T} + \mathbf{D}(\epsilon)\big)\mathbf{U}^* = \mathbf{A} + \mathbf{E}(\epsilon),$$

where $\mathbf{E}(\epsilon) = \mathbf{U}\mathbf{D}(\epsilon)\mathbf{U}^*$. The $(\lambda_i + \epsilon_i)$'s are the eigenvalues of $\mathbf{B}(\epsilon)$ and they are distinct, so $\mathbf{B}(\epsilon)$ is diagonalizable—by (7.2.6). Consequently, $\mathbf{B}(\epsilon)$ satisfies its own characteristic equation $0 = p_\epsilon(\lambda) = \det\,(\mathbf{A} + \mathbf{E}(\epsilon) - \lambda\mathbf{I})$ for each $\epsilon \neq 0$. The coefficients of $p_\epsilon(\lambda)$ are continuous functions of the entries in $\mathbf{E}(\epsilon)$ (recall (7.1.6)) and hence are continuous functions of the ϵ_i's. Combine this with $\lim_{\epsilon\to 0} \mathbf{E}(\epsilon) = \mathbf{0}$ to obtain $\mathbf{0} = \lim_{\epsilon\to 0} p_\epsilon(\mathbf{B}(\epsilon)) = p(\mathbf{A})$.

Note: Embedded in the above development is the fact that every square complex matrix is arbitrarily close to some diagonalizable matrix because for each $\epsilon \neq 0$, we have $\|\mathbf{A} - \mathbf{B}(\epsilon)\|_F = \|\mathbf{E}(\epsilon)\|_F = \epsilon$ (recall Exercise 5.6.9).

Example 7.3.7

Power method[74] is an iterative technique for computing a dominant eigenpair $(\lambda_1, \mathbf{x})$ of a diagonalizable $\mathbf{A} \in \Re^{m\times m}$ with eigenvalues

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_k|.$$

Note that this implies λ_1 is real—otherwise $\overline{\lambda}_1$ is another eigenvalue with the same magnitude as λ_1. Consider $f(z) = (z/\lambda_1)^n$, and use the spectral representation (7.3.6) along with $|\lambda_i/\lambda_1| < 1$ for $i = 2, 3, \ldots, k$ to conclude that

$$\begin{aligned}\left(\frac{\mathbf{A}}{\lambda_1}\right)^n &= f(\mathbf{A}) = f(\lambda_1)\mathbf{G}_1 + f(\lambda_2)\mathbf{G}_2 + \cdots + f(\lambda_k)\mathbf{G}_k \\ &= \mathbf{G}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^n \mathbf{G}_2 + \cdots + \left(\frac{\lambda_k}{\lambda_1}\right)^n \mathbf{G}_k \to \mathbf{G}_1\end{aligned} \tag{7.3.16}$$

as $n \to \infty$. Consequently, $(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n) \to \mathbf{G}_1\mathbf{x}_0 \in N\,(\mathbf{A} - \lambda_1\mathbf{I})$ for all $\mathbf{x}_0$. So if $\mathbf{G}_1\mathbf{x}_0 \neq \mathbf{0}$ or, equivalently, $\mathbf{x}_0 \notin R\,(\mathbf{A} - \lambda_1\mathbf{I})$, then $\mathbf{A}^n\mathbf{x}_0/\lambda_1^n$ converges to an eigenvector associated with λ_1. This means that the direction of $\mathbf{A}^n\mathbf{x}_0$ tends toward the direction of an eigenvector because λ_1^n acts only as a scaling factor to keep the length of $\mathbf{A}^n\mathbf{x}_0$ under control. Rather than using λ_1^n, we can scale $\mathbf{A}^n\mathbf{x}_0$ with something more convenient. For example, $\|\mathbf{A}^n\mathbf{x}_0\|$ (for any vector norm) is a reasonable scaling factor, but there are even better choices. For vectors $\mathbf{v}$, let $m(\mathbf{v})$ denote the component of maximal magnitude, and if there is more

[74] While the development of the power method was considered to be a great achievement when R. von Mises introduced it in 1929, later algorithms relegated its computational role to that of a special purpose technique. Nevertheless, it's still an important idea because, in some way or another, most practical algorithms for eigencomputations implicitly rely on the mathematical essence of the power method.

than one maximal component, let $m(\mathbf{v})$ be the *first* maximal component—e.g., $m(1,3,-2) = 3$, and $m(-3,3,-2) = -3$. It's clear that $m(\alpha\mathbf{v}) = \alpha m(\mathbf{v})$ for all scalars α. Suppose $m(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n) \to \gamma$. Since $(\mathbf{A}^n/\lambda_1^n) \to \mathbf{G}_1$, we see that

$$\lim_{n\to\infty} \frac{\mathbf{A}^n\mathbf{x}_0}{m(\mathbf{A}^n\mathbf{x}_0)} = \lim_{n\to\infty} \frac{(\mathbf{A}^n/\lambda_1^n)\mathbf{x}_0}{m(\mathbf{A}^n\mathbf{x}_0/\lambda_1^n)} = \frac{\mathbf{G}_1\mathbf{x}_0}{\gamma} = \mathbf{x}$$

is an eigenvector associated with λ_1. But rather than successively powering $\mathbf{A}$, the sequence $\mathbf{A}^n\mathbf{x}_0/m(\mathbf{A}^n\mathbf{x}_0)$ is more efficiently generated by starting with $\mathbf{x}_0 \notin R(\mathbf{A} - \lambda_1\mathbf{I})$ and setting

$$\mathbf{y}_n = \mathbf{A}\mathbf{x}_n, \quad \nu_n = m(\mathbf{y}_n), \quad \mathbf{x}_{n+1} = \frac{\mathbf{y}_n}{\nu_n}, \quad \text{for } n = 0, 1, 2, \ldots. \tag{7.3.17}$$

Not only does $\mathbf{x}_n \to \mathbf{x}$, but as a bonus we get $\nu_n \to \lambda_1$ because for all n, $\mathbf{A}\mathbf{x}_{n+1} = \mathbf{A}^2\mathbf{x}_n/\nu_n$, so if $\nu_n \to \nu$ as $n \to \infty$, the limit on the left-hand side is $\mathbf{A}\mathbf{x} = \lambda_1\mathbf{x}$, while the limit on the right-hand side is $\mathbf{A}^2\mathbf{x}/\nu = \lambda_1^2\mathbf{x}/\nu$. Since these two limits must agree, $\lambda_1\mathbf{x} = (\lambda_1^2/\nu)\mathbf{x}$, and this implies $\nu = \lambda_1$.

Summary. The sequence $(\nu_n, \mathbf{x}_n)$ defined by (7.3.17) converges to an eigenpair $(\lambda_1, \mathbf{x})$ for $\mathbf{A}$ provided that $\mathbf{G}_1\mathbf{x}_0 \neq \mathbf{0}$ or, equivalently, $\mathbf{x}_0 \notin R(\mathbf{A} - \lambda_1\mathbf{I})$.

- ▷ **Advantages.** Each iteration requires only one matrix–vector product, and this can be exploited to reduce the computational effort when $\mathbf{A}$ is large and sparse—assuming that a dominant eigenpair is the only one of interest.
- ▷ **Disadvantages.** Only a dominant eigenpair is determined—something else must be done if others are desired. Furthermore, it's clear from (7.3.16) that the rate at which (7.3.17) converges depends on how fast $(\lambda_2/\lambda_1)^n \to 0$, so convergence is slow when $|\lambda_1|$ is close to $|\lambda_2|$.

Example 7.3.8

Inverse Power Method. Given a real approximation $\alpha \notin \sigma(\mathbf{A})$ to any real $\lambda \in \sigma(\mathbf{A})$, this algorithm (also called the ***inverse iteration***) determines an eigenpair $(\lambda, \mathbf{x})$ for a diagonalizable matrix $\mathbf{A} \in \Re^{m\times m}$ by applying the power method[75] to $\mathbf{B} = (\mathbf{A} - \alpha\mathbf{I})^{-1}$. Recall from Exercise 7.1.9 that

$$\begin{gathered} \mathbf{x} \text{ is an eigenvector for } \mathbf{A} \iff \mathbf{x} \text{ is an eigenvector for } \mathbf{B}, \\ \lambda \in \sigma(\mathbf{A}) \iff (\lambda - \alpha)^{-1} \in \sigma(\mathbf{B}). \end{gathered} \tag{7.3.18}$$

If $|\lambda - \alpha| < |\lambda_i - \alpha|$ for all other $\lambda_i \in \sigma(\mathbf{A})$, then $(\lambda - \alpha)^{-1}$ is the dominant eigenvalue of $\mathbf{B}$ because $|\lambda - \alpha|^{-1} > |\lambda_i - \alpha|^{-1}$. Therefore, applying the power

[75] The relation between the power method and inverse iteration is clear to us now, but it originally took 15 years to make the connection. Inverse iteration was not introduced until 1944 by the German mathematician Helmut Wielandt (1910–2001).

method to $\mathbf{B}$ produces an eigenpair $\big((\lambda-\alpha)^{-1},\mathbf{x}\big)$ for $\mathbf{B}$ from which the eigenpair $(\lambda,\mathbf{x})$ for $\mathbf{A}$ is determined. That is, if $\mathbf{x}_0 \notin R(\mathbf{B}-\lambda\mathbf{I})$, and if

$$\mathbf{y}_n = \mathbf{B}\mathbf{x}_n = (\mathbf{A}-\alpha\mathbf{I})^{-1}\mathbf{x}_n, \quad \nu_n = m(\mathbf{y}_n), \quad \mathbf{x}_{n+1} = \frac{\mathbf{y}_n}{\nu_n} \quad \text{for } n = 0,1,2,\ldots,$$

then $(\nu_n,\mathbf{x}_n) \to \big((\lambda-\alpha)^{-1},\mathbf{x}\big)$, an eigenpair for $\mathbf{B}$, so (7.3.18) guarantees that $(\nu_n^{-1}+\alpha,\ \mathbf{x}_n) \to (\lambda,\mathbf{x})$, an eigenpair for $\mathbf{A}$. Rather than using matrix inversion to compute $\mathbf{y}_n = (\mathbf{A}-\alpha\mathbf{I})^{-1}\mathbf{x}_n$, it's more efficient to solve the linear system $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ for $\mathbf{y}_n$. Because this is a system in which the coefficient matrix remains the same from step to step, the efficiency is further enhanced by computing an LU factorization of $(\mathbf{A}-\alpha\mathbf{I})$ at the outset so that at each step only one forward solve and one back solve (as described on pp. 146 and 153) are needed to determine $\mathbf{y}_n$.

- ▷ **Advantages.** Striking results are often obtained (particularly in the case of symmetric matrices) with only one or two iterations, even when $\mathbf{x}_0$ is nearly in $R(\mathbf{B}-\lambda\mathbf{I}) = R(\mathbf{A}-\lambda\mathbf{I})$. For α close to λ, computing an accurate floating-point solution of $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ is difficult because $\mathbf{A}-\alpha\mathbf{I}$ is nearly singular, and this almost surely guarantees that $(\mathbf{A}-\alpha\mathbf{I})\mathbf{y}_n = \mathbf{x}_n$ is an ill-conditioned system. But only the direction of the solution is important, and the direction of a computed solution is usually reasonable in spite of conditioning problems. Finally, the algorithm can be adapted to compute approximations of eigenvectors associated with complex eigenvalues.
- ▷ **Disadvantages.** Only one eigenpair at a time is computed, and an approximate eigenvalue must be known in advance. Furthermore, the rate of convergence depends on how fast $[(\lambda-\alpha)/(\lambda_i-\alpha)]^n \to 0$, and this can be slow when there is another eigenvalue λ_i close to the desired λ. If λ_i is too close to λ, roundoff error can divert inverse iteration toward an eigenvector associated with λ_i instead of λ in spite of a theoretically correct α.

Note: In the standard version of inverse iteration a constant value of α is used at each step to approximate an eigenvalue λ, but there is variation called ***Rayleigh quotient iteration*** that uses the current iterate $\mathbf{x}_n$ to improve the value of α at each step by setting $\alpha = \mathbf{x}_n^T\mathbf{A}\mathbf{x}_n/\mathbf{x}_n^T\mathbf{x}_n$. The function $R(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}/\mathbf{x}^T\mathbf{x}$ is called the *Rayleigh quotient.* It can be shown that if $\mathbf{x}$ is a good approximation to an eigenvector, then $R(\mathbf{x})$ is a good approximation of the associated eigenvalue. More is said about this in Example 7.5.1 (p. 549).

Example 7.3.9

The QR Iteration algorithm for computing the eigenvalues of a general matrix came from an elegantly simple idea that was proposed by Heinz Rutishauser in 1958 and refined by J. F. G. Francis in 1961-1962. The underlying concept is to alternate between computing QR factors (Rutishauser used LU factors) and

reversing their order as shown below. Starting with $\mathbf{A}_1 = \mathbf{A} \in \Re^{n\times n}$,

$$\begin{aligned}
&\text{Factor:} && \mathbf{A}_1 = \mathbf{Q}_1\mathbf{R}_1,\\
&\text{Set:} && \mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1,\\
&\text{Factor:} && \mathbf{A}_2 = \mathbf{Q}_2\mathbf{R}_2,\\
&\text{Set:} && \mathbf{A}_3 = \mathbf{R}_2\mathbf{Q}_2,\\
& && \vdots
\end{aligned}$$

In general, $\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k$, where $\mathbf{Q}_k$ and $\mathbf{R}_k$ are the QR factors of $\mathbf{A}_k$. Notice that if $\mathbf{P}_k = \mathbf{Q}_1\mathbf{Q}_2\cdots\mathbf{Q}_k$, then each $\mathbf{P}_k$ is an orthogonal matrix such that

$$\begin{aligned}
\mathbf{P}_1^T\mathbf{A}\mathbf{P}_1 &= \mathbf{Q}_1^T\mathbf{Q}_1\mathbf{R}_1\mathbf{Q}_1 = \mathbf{A}_2,\\
\mathbf{P}_2^T\mathbf{A}\mathbf{P}_2 &= \mathbf{Q}_2^T\mathbf{Q}_1^T\mathbf{A}\mathbf{Q}_1\mathbf{Q}_2 = \mathbf{Q}_2^T\mathbf{A}_2\mathbf{Q}_2 = \mathbf{A}_3,\\
&\vdots\\
\mathbf{P}_k^T\mathbf{A}\mathbf{P}_k &= \mathbf{A}_{k+1}.
\end{aligned}$$

In other words, $\mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4, \ldots$ are each orthogonally similar to $\mathbf{A}$, and hence $\sigma(\mathbf{A}_k) = \sigma(\mathbf{A})$ for each k. But the process does more than just create a matrix that is similar to $\mathbf{A}$ at each step. The magic lies in the fact that if the process converges, then $\lim_{k\to\infty}\mathbf{A}_k = \mathbf{R}$ is an upper-triangular matrix in which the diagonal entries are the eigenvalues of $\mathbf{A}$. Indeed, if $\mathbf{P}_k \to \mathbf{P}$, then

$$\mathbf{Q}_k = \mathbf{P}_{k-1}^T\mathbf{P}_k \to \mathbf{P}^T\mathbf{P} = \mathbf{I} \quad \text{and} \quad \mathbf{R}_k = \mathbf{A}_{k+1}\mathbf{Q}_k^T \to \mathbf{R}\,\mathbf{I} = \mathbf{R},$$

so

$$\lim_{k\to\infty}\mathbf{A}_k = \lim_{k\to\infty}\mathbf{Q}_k\mathbf{R}_k = \mathbf{R},$$

which is necessarily upper triangular having diagonal entries equal to the eigenvalues of $\mathbf{A}$. However, as is often the case, there is a big gap between theory and practice, and turning this clever idea into a practical algorithm requires significant effort. For example, one obvious hurdle that needs to be overcome is the fact that the R factor in a QR factorization has positive diagonal entries, so, unless modifications are made, the "vanilla" version of the QR iteration can't converge for matrices with complex or nonpositive eigenvalues. Laying out all of the details and analyzing the rigors that constitute the practical implementation of the QR iteration is tedious and would take us too far astray, but the basic principals are within our reach.

- **Hessenberg Matrices.** A big step in turning the QR iteration into a practical method is to realize that everything can be done with upper-Hessenberg matrices. As discussed in Example 5.7.4 (p. 350), Householder reduction can be used to produce an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{H}_1$, and Example 5.7.5 (p. 352) shows that Givens reduction easily produces

the QR factors of any Hessenberg matrix. Givens reduction on $\mathbf{H}_1$ produces the Q factor of $\mathbf{H}_1$ as the transposed product of plane rotations $\mathbf{Q}_1 = \mathbf{P}_{12}^T\mathbf{P}_{23}^T\cdots\mathbf{P}_{(n-1)n}^T$, and this is also upper Hessenberg (constructing a 4×4 example will convince you). Since multiplication by an upper-triangular matrix can't alter the upper-Hessenberg structure, the matrix $\mathbf{R}_1\mathbf{Q}_1 = \mathbf{H}_2$ at the second step of the QR iteration is again upper Hessenberg, and so on for each successive step. Being able to iterate with Hessenberg matrices results in a significant reduction of arithmetic. Note that if $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{H}_k = \mathbf{H}_k^T$ for each k, which means that each $\mathbf{H}_k$ is tridiagonal in structure.

- **Convergence.** When the $\mathbf{H}_k$'s converge, the entries at the bottom of the first subdiagonal tend to die first—i.e., a typical pattern might be

$$\mathbf{H}_k = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \epsilon & \star \end{pmatrix}.$$

When ϵ is satisfactorily small, take $\star$ (the (n,n)-entry) to be an eigenvalue, and deflate the problem. An even nicer state of affairs is to have a zero (or a satisfactorily small) entry in row $n-1$ and column 2 (illustrated below for $n=4$)

$$\mathbf{H}_k = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & \epsilon & \star & \star \\ 0 & 0 & \star & \star \end{pmatrix} \tag{7.3.19}$$

because the trailing 2×2 block $\begin{pmatrix} \star & \star \\ \star & \star \end{pmatrix}$ will yield two eigenvalues by the quadratic formula, and thus complex eigenvalues can be revealed.

- **Shifts.** Instead of factoring $\mathbf{H}_k$ at the kth step, factor a shifted matrix $\mathbf{H}_k - \alpha_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k$, and set $\mathbf{H}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \alpha_k\mathbf{I}$, where α_k is an approximate *real* eigenvalue—a good candidate is $\alpha_k = [\mathbf{H}_k]_{nn}$. Notice that $\sigma(\mathbf{H}_{k+1}) = \sigma(\mathbf{H}_k)$ because $\mathbf{H}_{k+1} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k$. The inverse power method is now at work. To see how, drop the subscripts, and write $\mathbf{H} - \alpha\mathbf{I} = \mathbf{QR}$ as $\mathbf{Q}^T = \mathbf{R}(\mathbf{H}-\alpha\mathbf{I})^{-1}$. If $\alpha \approx \lambda \in \sigma(\mathbf{H}) = \sigma(\mathbf{A})$ (say, $|\lambda - \alpha| = \epsilon$ with $\alpha, \lambda \in \Re$), then the discussion concerning the inverse power method in Example 7.3.8 insures that the rows in $\mathbf{Q}^T$ are close to being left-hand eigenvectors of $\mathbf{H}$ associated with λ. In particular, if $\mathbf{q}_n^T$ is the last row in $\mathbf{Q}^T$, then

$$r_{nn}\mathbf{e}_n^T = \mathbf{e}_n^T\mathbf{R} = \mathbf{q}_n^T\mathbf{QR} = \mathbf{q}_n^T(\mathbf{H}-\alpha\mathbf{I}) = \mathbf{q}_n^T\mathbf{H} - \alpha\mathbf{q}_n^T \approx (\lambda-\alpha)\mathbf{q}_n^T,$$

so $r_{nn} = |r_{nn}| \approx \left\|(\lambda-\alpha)\mathbf{q}_n^T\right\|_2 = \epsilon$ and $\mathbf{q}_n^T \approx \pm\mathbf{e}_n^T$. The significance of this

is revealed by looking at a generic 4×4 pattern for

$$\begin{aligned}
\mathbf{H}_{k+1} &= \mathbf{RQ} + \alpha\mathbf{I} \\
&\approx \begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & \epsilon \end{pmatrix} \begin{pmatrix} * & * & * & 0 \\ * & * & * & 0 \\ 0 & * & * & 0 \\ 0 & 0 & \star & \pm 1 \end{pmatrix} + \begin{pmatrix} \alpha & & & \\ & \alpha & & \\ & & \alpha & \\ & & & \alpha \end{pmatrix} \\
&= \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \epsilon\star & \alpha \pm \epsilon \end{pmatrix} \approx \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & 0 & \alpha \pm \epsilon \end{pmatrix}.
\end{aligned}$$

The strength of the last approximation rests not only on the size of ϵ, but it is also reinforced by the fact that $\star \approx 0$ because the 2-norm of the last row of $\mathbf{Q}$ must be 1. This indicates why this technique (called the *single shifted* QR iteration) can provide rapid convergence to a real eigenvalue. To extract complex eigenvalues, a *double shift* strategy is employed in which the eigenvalues α_k and β_k of the lower 2×2 block of $\mathbf{H}_k$ are used as shifts as indicated below:

$$\begin{aligned}
&\text{Factor:} && \mathbf{H}_k - \alpha_k\mathbf{I} = \mathbf{Q}_k\mathbf{R}_k, \\
&\text{Set:} && \mathbf{H}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + \alpha_k\mathbf{I} && (\text{so } \mathbf{H}_{k+1} = \mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k), \\
&\text{Factor:} && \mathbf{H}_{k+1} - \beta_k\mathbf{I} = \mathbf{Q}_{k+1}\mathbf{R}_{k+1}, \\
&\text{Set:} && \mathbf{H}_{k+2} = \mathbf{R}_{k+1}\mathbf{Q}_{k+1} + \beta_k\mathbf{I} && (\text{so } \mathbf{H}_{k+2} = \mathbf{Q}_{k+1}^T\mathbf{Q}_k^T\mathbf{H}_k\mathbf{Q}_k\mathbf{Q}_{k+1}), \\
& && \vdots
\end{aligned}$$

The nice thing about the double shift strategy is that even when α_k is complex (so that $\beta_k = \overline{\alpha}_k$) the matrix $\mathbf{Q}_k\mathbf{Q}_{k+1}$ (and hence $\mathbf{H}_{k+2}$) is real, and there are efficient ways to form $\mathbf{Q}_k\mathbf{Q}_{k+1}$ by computing only the first column of the product. The double shift method typically requires very few iterations (using only real arithmetic) to produce a small entry in the $(n-2, 2)$-position as depicted in (7.3.19) for a generic 4×4 pattern.

Exercises for section 7.3

7.3.1. Determine $\cos \mathbf{A}$ for $\mathbf{A} = \begin{pmatrix} -\pi/2 & \pi/2 \\ \pi/2 & -\pi/2 \end{pmatrix}$.

7.3.2. For the matrix $\mathbf{A}$ in Example 7.3.3, verify with direct computation that $e^{\lambda_1 t}\mathbf{G}_1 + e^{\lambda_2 t}\mathbf{G}_2 = \mathbf{P}\begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}\mathbf{P}^{-1} = e^{\mathbf{A}t}$.

7.3.3. Explain why $\sin^2 \mathbf{A} + \cos^2 \mathbf{A} = \mathbf{I}$ for a diagonalizable matrix $\mathbf{A}$.

7.3.4. Explain $e^{\mathbf{0}} = \mathbf{I}$ for every square zero matrix.

7.3.5. The ***spectral mapping property*** for diagonalizable matrices says that if $f(\mathbf{A})$ exists, and if $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues of $\mathbf{A}_{n\times n}$ (including multiplicities), then $\{f(\lambda_1), \ldots, f(\lambda_n)\}$ are the eigenvalues of $f(\mathbf{A})$.

(a) Establish this for diagonalizable matrices.

(b) Establish this when an infinite series $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ defines $f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n$ as discussed in (7.3.7).

7.3.6. Explain why $\det\left(e^{\mathbf{A}}\right) = e^{trace(\mathbf{A})}$.

7.3.7. Suppose that for nondiagonalizable matrices $\mathbf{A}_{m\times m}$ an infinite series $f(z) = \sum_{n=0}^{\infty} c_n(z - z_0)^n$ is used to define $f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n(\mathbf{A} - z_0\mathbf{I})^n$ as suggested in (7.3.7). Neglecting convergence issues, explain why there is a polynomial $p(z)$ of at most degree $m-1$ such that $f(\mathbf{A}) = p(\mathbf{A})$.

7.3.8. If $f(\mathbf{A})$ exists for a diagonalizable $\mathbf{A}$, explain why $\mathbf{A}f(\mathbf{A}) = f(\mathbf{A})\mathbf{A}$. What can you say when $\mathbf{A}$ is not diagonalizable?

7.3.9. Explain why $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ whenever $\mathbf{AB} = \mathbf{BA}$. Give an example to show that $e^{\mathbf{A}+\mathbf{B}}$, $e^{\mathbf{A}}e^{\mathbf{B}}$, and $e^{\mathbf{B}}e^{\mathbf{A}}$ all can differ when $\mathbf{AB} \neq \mathbf{BA}$. **Hint:** Exercise 7.2.16 can be used for the diagonalizable case. For the general case, consider $\mathbf{F}(t) = e^{(\mathbf{A}+\mathbf{B})t} - e^{\mathbf{A}t}e^{\mathbf{B}t}$ and $\mathbf{F}'(t)$.

7.3.10. Show that $e^{\mathbf{A}}$ is an orthogonal matrix whenever $\mathbf{A}$ is skew symmetric.

7.3.11. A particular electronic device consists of a collection of switching circuits that can be either in an ON state or an OFF state. These electronic switches are allowed to change state at regular time intervals called *clock cycles.* Suppose that at the end of each clock cycle, 30% of the switches currently in the OFF state change to ON, while 90% of those in the ON state revert to the OFF state.

(a) Show that the device approaches an equilibrium in the sense that the proportion of switches in each state eventually becomes constant, and determine these equilibrium proportions.

(b) Independent of the initial proportions, about how many clock cycles does it take for the device to become essentially stable?

7.3.12. The ***spectral radius*** of $\mathbf{A}$ is $\rho(\mathbf{A}) = \max_{\lambda_i \in \sigma(\mathbf{A})} |\lambda_i|$ (p. 497). Prove that if $\mathbf{A}$ is diagonalizable, then

$$\rho(\mathbf{A}) = \lim_{n\to\infty} \|\mathbf{A}^n\|^{1/n} \quad \text{for every matrix norm.}$$

This result is true for nondiagonalizable matrices as well, but the proof at this point in the game is more involved. The full development is given in Example 7.10.1 (p. 619).

7.3.13. Find a dominant eigenpair for $\mathbf{A} = \begin{pmatrix} 7 & 2 & 3 \\ 0 & 2 & 0 \\ -6 & -2 & -2 \end{pmatrix}$ by the power method.

7.3.14. Apply the inverse power method (Example 7.3.8, p. 534) to find an eigenvector for each of the eigenvalues of the matrix $\mathbf{A}$ in Exercise 7.3.13.

7.3.15. Explain why the function $m(\mathbf{v})$ used in the development of the power method in Example 7.3.7 is not a continuous function, so statements like $m(\mathbf{x}_n) \to m(\mathbf{x})$ when $\mathbf{x}_n \to \mathbf{x}$ are not valid. Nevertheless, if $\lim_{n\to\infty} \mathbf{x}_n \neq \mathbf{0}$, then $\lim_{n\to\infty} m(\mathbf{x}_n) \neq 0$.

7.3.16. Let $\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & -2 & -1 \\ 0 & 2 & 1 \end{pmatrix}$.

(a) Apply the "vanilla" QR iteration to $\mathbf{H}$.
(b) Apply the the single shift QR iteration on $\mathbf{H}$.

7.3.17. Show that the QR iteration can fail to converge using $\mathbf{H} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

(a) First use the "vanilla" QR iteration on $\mathbf{H}$ to see what happens.
(b) Now try the single shift QR iteration on $\mathbf{H}$.
(c) Finally, execute the double shift QR iteration on $\mathbf{H}$.

7.4 SYSTEMS OF DIFFERENTIAL EQUATIONS

Systems of first-order linear differential equations with constant coefficients were used in §7.1 to motivate the introduction of eigenvalues and eigenvectors, but now we can delve a little deeper. For constants a_{ij}, the goal is to solve the following system for the unknown functions $u_i(t)$.

$$\begin{array}{lcl} u_1' = a_{11}u_1 + a_{12}u_2 + \cdots + a_{1n}u_n, & & u_1(0) = c_1, \\ u_2' = a_{21}u_1 + a_{22}u_2 + \cdots + a_{2n}u_n, & & u_2(0) = c_2, \\ \quad\vdots & \text{with} & \quad\vdots \\ u_n' = a_{n1}u_1 + a_{n2}u_2 + \cdots + a_{nn}u_n, & & u_n(0) = c_n. \end{array} \tag{7.4.1}$$

Since the scalar exponential provides the unique solution to a single differential equation $u'(t) = \alpha u(t)$ with $u(0) = c$ as $u(t) = \mathrm{e}^{\alpha t}c$, it's only natural to try to use the matrix exponential in an analogous way to solve a system of differential equations. Begin by writing (7.4.1) in matrix form as $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where

$$\mathbf{u} = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_n(t) \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$

If $\mathbf{A}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then (7.3.6) guarantees

$$\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\lambda_1 t}\mathbf{G}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{G}_2 + \cdots + \mathrm{e}^{\lambda_k t}\mathbf{G}_k. \tag{7.4.2}$$

The following identities are derived from properties of the $\mathbf{G}_i$'s given on p. 517.

- $d\mathrm{e}^{\mathbf{A}t}/dt = \sum_{i=1}^k \lambda_i \mathrm{e}^{\lambda_i t}\mathbf{G}_i = \left(\sum_{i=1}^k \lambda_i \mathbf{G}_i\right)\left(\sum_{i=1}^k \mathrm{e}^{\lambda_i t}\mathbf{G}_i\right) = \mathbf{A}\mathrm{e}^{\mathbf{A}t}$. (7.4.3)
- $\mathbf{A}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathbf{A}$ (by a similar argument). (7.4.4)
- $\mathrm{e}^{-\mathbf{A}t}\mathrm{e}^{\mathbf{A}t} = \mathrm{e}^{\mathbf{A}t}\mathrm{e}^{-\mathbf{A}t} = \mathbf{I} = \mathrm{e}^{\mathbf{0}}$ (by a similar argument). (7.4.5)

Equation (7.4.3) insures that $\mathbf{u} = \mathrm{e}^{\mathbf{A}t}\mathbf{c}$ is *one* solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. To see that $\mathbf{u} = \mathrm{e}^{\mathbf{A}t}\mathbf{c}$ is the *only* solution, suppose $\mathbf{v}(t)$ is another solution so that $\mathbf{v}' = \mathbf{A}\mathbf{v}$ with $\mathbf{v}(0) = \mathbf{c}$. Differentiating $\mathrm{e}^{-\mathbf{A}t}\mathbf{v}$ produces

$$\frac{d\left[\mathrm{e}^{-\mathbf{A}t}\mathbf{v}\right]}{dt} = \mathrm{e}^{-\mathbf{A}t}\mathbf{v}' - \mathrm{e}^{-\mathbf{A}t}\mathbf{A}\mathbf{v} = \mathbf{0}, \quad \text{so} \quad \mathrm{e}^{-\mathbf{A}t}\mathbf{v} \text{ is constant for all } t.$$

At $t = 0$ we have $e^{-\mathbf{A}t}\mathbf{v}\big|_{t=0} = e^{\mathbf{0}}\mathbf{v}(0) = \mathbf{I}\mathbf{c} = \mathbf{c}$, and hence $e^{-\mathbf{A}t}\mathbf{v} = \mathbf{c}$ for all t. Multiply both sides of this equation by $e^{\mathbf{A}t}$ and use (7.4.5) to conclude $\mathbf{v} = e^{\mathbf{A}t}\mathbf{c}$. Thus $\mathbf{u} = e^{\mathbf{A}t}\mathbf{c}$ is the unique solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$ with $\mathbf{u}(0) = \mathbf{c}$.

Finally, notice that $\mathbf{v}_i = \mathbf{G}_i\mathbf{c} \in N(\mathbf{A} - \lambda_i\mathbf{I})$ is an eigenvector associated with λ_i, so that the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, is

$$\mathbf{u} = e^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_2 t}\mathbf{v}_2 + \cdots + e^{\lambda_k t}\mathbf{v}_k, \tag{7.4.6}$$

and this solution is completely determined by the eigenpairs $(\lambda_i, \mathbf{v}_i)$. It turns out that $\mathbf{u}$ also can be expanded in terms of *any* complete set of independent eigenvectors—see Exercise 7.4.1. Let's summarize what's been said so far.

Differential Equations

If $\mathbf{A}_{n\times n}$ is diagonalizable with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then the unique solution of $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, is given by

$$\mathbf{u} = e^{\mathbf{A}t}\mathbf{c} = e^{\lambda_1 t}\mathbf{v}_1 + e^{\lambda_2 t}\mathbf{v}_2 + \cdots + e^{\lambda_k t}\mathbf{v}_k \tag{7.4.7}$$

in which $\mathbf{v}_i$ is the eigenvector $\mathbf{v}_i = \mathbf{G}_i\mathbf{c}$, where $\mathbf{G}_i$ is the ith spectral projector. (See Exercise 7.4.1 for an alternate eigenexpansion.) Nonhomogeneous systems as well as the nondiagonalizable case are treated in Example 7.9.6 (p. 608).

Example 7.4.1

An Application to Diffusion. Important issues in medicine and biology involve the question of how drugs or chemical compounds move from one cell to another by means of diffusion through cell walls. Consider two cells, as depicted in Figure 7.4.1, which are both devoid of a particular compound. A unit amount of the compound is injected into the first cell at time $t = 0$, and as time proceeds the compound diffuses according to the following assumption.

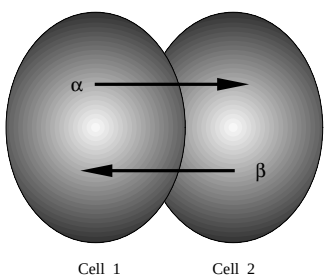


Figure 7.4.1

At each point in time the rate (amount per second) of diffusion from one cell to the other is proportional to the concentration (amount per unit volume) of the compound in the cell giving up the compound—say the rate of diffusion from cell 1 to cell 2 is α times the concentration in cell 1, and the rate of diffusion from cell 2 to cell 1 is β times the concentration in cell 2. Assume $\alpha,\ \beta > 0$.

Problem: Determine the concentration of the compound in each cell at any given time t, and, in the long run, determine the steady-state concentrations.

Solution: If $u_k = u_k(t)$ denotes the concentration of the compound in cell k at time t, then the statements in the above assumption are translated as follows:

$$\begin{aligned}
\frac{du_1}{dt} &= \text{rate in} - \text{rate out} = \beta u_2 - \alpha u_1, \quad \text{where} \quad u_1(0) = 1,\\
\frac{du_2}{dt} &= \text{rate in} - \text{rate out} = \alpha u_1 - \beta u_2, \quad \text{where} \quad u_2(0) = 0.
\end{aligned}$$

In matrix notation this system is $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where

$$\mathbf{A} = \begin{pmatrix} -\alpha & \beta \\ \alpha & -\beta \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Since $\mathbf{A}$ is the matrix of Example 7.3.3 we can use the results from Example 7.3.3 to write the solution as

$$\mathbf{u}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c} = \frac{1}{\alpha+\beta}\left[\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} + \mathrm{e}^{-(\alpha+\beta)t}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix}\right]\begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

so that

$$u_1(t) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\mathrm{e}^{-(\alpha+\beta)t} \quad \text{and} \quad u_2(t) = \frac{\alpha}{\alpha+\beta}\left(1 - \mathrm{e}^{-(\alpha+\beta)t}\right).$$

In the long run, the concentrations in each cell stabilize in the sense that

$$\lim_{t\to\infty} u_1(t) = \frac{\beta}{\alpha+\beta} \quad \text{and} \quad \lim_{t\to\infty} u_2(t) = \frac{\alpha}{\alpha+\beta}.$$

An innumerable variety of physical situations can be modeled by $\mathbf{u}' = \mathbf{A}\mathbf{u}$, and the form of the solution (7.4.6) makes it clear that the eigenvalues and eigenvectors of $\mathbf{A}$ are intrinsic to the underlying physical phenomenon being investigated. We might say that the eigenvalues and eigenvectors of $\mathbf{A}$ act as its genes and chromosomes because they are the basic components that either dictate or govern all other characteristics of $\mathbf{A}$ along with the physics of associated phenomena.

For example, consider the long-run behavior of a physical system that can be modeled by $\mathbf{u}' = \mathbf{A}\mathbf{u}$. We usually want to know whether the system will eventually blow up or will settle down to some sort of stable state. Might it neither blow up nor settle down but rather oscillate indefinitely? These are questions concerning the nature of the limit

$$\lim_{t\to\infty} \mathbf{u}(t) = \lim_{t\to\infty} \mathrm{e}^{\mathbf{A}t}\mathbf{c} = \lim_{t\to\infty} \left(\mathrm{e}^{\lambda_1 t}\mathbf{G}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{G}_2 + \cdots + \mathrm{e}^{\lambda_k t}\mathbf{G}_k\right)\mathbf{c},$$

and the answers depend only on the eigenvalues. To see how, recall that for a complex number $\lambda = x + \mathrm{i}y$ and a real parameter $t > 0$,

$$\mathrm{e}^{\lambda t} = \mathrm{e}^{(x+\mathrm{i}y)t} = \mathrm{e}^{xt}\mathrm{e}^{\mathrm{i}yt} = \mathrm{e}^{xt}\left(\cos yt + \mathrm{i}\sin yt\right). \tag{7.4.8}$$

The term $\mathrm{e}^{\mathrm{i}yt} = (\cos yt + \mathrm{i}\sin yt)$ is a point on the unit circle that oscillates as a function of t, so $|\mathrm{e}^{\mathrm{i}yt}| = |\cos yt + \mathrm{i}\sin yt| = 1$ and $\left|\mathrm{e}^{\lambda t}\right| = |\mathrm{e}^{xt}\mathrm{e}^{\mathrm{i}yt}| = |\mathrm{e}^{xt}| = \mathrm{e}^{xt}$. This makes it clear that if $\mathrm{Re}\,(\lambda_i) < 0$ for each i, then, as $t \to \infty$, $\mathrm{e}^{\mathbf{A}t} \to \mathbf{0}$, and $\mathbf{u}(t) \to \mathbf{0}$ for every initial vector $\mathbf{c}$. Thus the system eventually settles down to zero, and we say the system is *stable.* On the other hand, if $\mathrm{Re}\,(\lambda_i) > 0$ for some i, then components of $\mathbf{u}(t)$ may become unbounded as $t \to \infty$, and we say the system is *unstable.* Finally, if $\mathrm{Re}\,(\lambda_i) \le 0$ for each i, then the components of $\mathbf{u}(t)$ remain finite for all t, but some may oscillate indefinitely, and this is called a *semistable* situation. Below is a summary of stability.

Stability

Let $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where $\mathbf{A}$ is diagonalizable with eigenvalues λ_i.

- If $\mathrm{Re}\,(\lambda_i) < 0$ for each i, then $\lim\limits_{t\to\infty} \mathrm{e}^{\mathbf{A}t} = \mathbf{0}$, and $\lim\limits_{t\to\infty} \mathbf{u}(t) = \mathbf{0}$ for every initial vector $\mathbf{c}$. In this case $\mathbf{u}' = \mathbf{A}\mathbf{u}$ is said to be a ***stable system,*** and $\mathbf{A}$ is called a ***stable matrix.***
- If $\mathrm{Re}\,(\lambda_i) > 0$ for some i, then components of $\mathbf{u}(t)$ can become unbounded as $t \to \infty$, in which case the system $\mathbf{u}' = \mathbf{A}\mathbf{u}$ as well as the underlying matrix $\mathbf{A}$ are said to be ***unstable.***
- If $\mathrm{Re}\,(\lambda_i) \le 0$ for each i, then the components of $\mathbf{u}(t)$ remain finite for all t, but some can oscillate indefinitely. This is called a ***semistable*** situation.

Example 7.4.2

Predator–Prey Application. Consider two species of which one is the predator and the other is the prey, and assume there are initially 100 in each population. Let $u_1(t)$ and $u_2(t)$ denote the respective population of the predator and

prey species at time t, and suppose their growth rates are given by

$$u_1' = u_1 + u_2,$$
$$u_2' = -u_1 + u_2.$$

Problem: Determine the size of each population at all future times, and decide if (and when) either population will eventually become extinct.

Solution: Write the system as $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} 100 \\ 100 \end{pmatrix}.$$

The characteristic equation for $\mathbf{A}$ is $p(\lambda) = \lambda^2 - 2\lambda + 2 = 0$, so the eigenvalues for $\mathbf{A}$ are $\lambda_1 = 1 + \mathrm{i}$ and $\lambda_2 = 1 - \mathrm{i}$. We know from (7.4.7) that

$$\mathbf{u}(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 \quad (\text{where } \mathbf{v}_i = \mathbf{G}_i\mathbf{c}\,) \tag{7.4.9}$$

is the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. The spectral theorem on p. 517 implies $\mathbf{A} - \lambda_2\mathbf{I} = (\lambda_1 - \lambda_2)\mathbf{G}_1$ and $\mathbf{I} = \mathbf{G}_1 + \mathbf{G}_2$, so $(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c} = (\lambda_1 - \lambda_2)\mathbf{v}_1$ and $\mathbf{c} = \mathbf{v}_1 + \mathbf{v}_2$, and consequently

$$\mathbf{v}_1 = \frac{(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c}}{(\lambda_1 - \lambda_2)} = 50\begin{pmatrix} \lambda_2 \\ \lambda_1 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \mathbf{c} - \mathbf{v}_1 = 50\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}.$$

With the aid of (7.4.8) we obtain the solution components from (7.4.9) as

$$u_1(t) = 50\left(\lambda_2\mathrm{e}^{\lambda_1 t} + \lambda_1\mathrm{e}^{\lambda_2 t}\right) = 100\mathrm{e}^t(\cos t + \sin t)$$

and

$$u_2(t) = 50\left(\lambda_1\mathrm{e}^{\lambda_1 t} + \lambda_2\mathrm{e}^{\lambda_2 t}\right) = 100\mathrm{e}^t(\cos t - \sin t).$$

The system is unstable because $\mathrm{Re}\,(\lambda_i) > 0$ for each eigenvalue. Indeed, $u_1(t)$ and $u_2(t)$ both become unbounded as $t \to \infty$. However, a population cannot become negative–once it's zero, it's extinct. Figure 7.4.2 shows that the graph of $u_2(t)$ will cross the horizontal axis before that of $u_1(t)$.

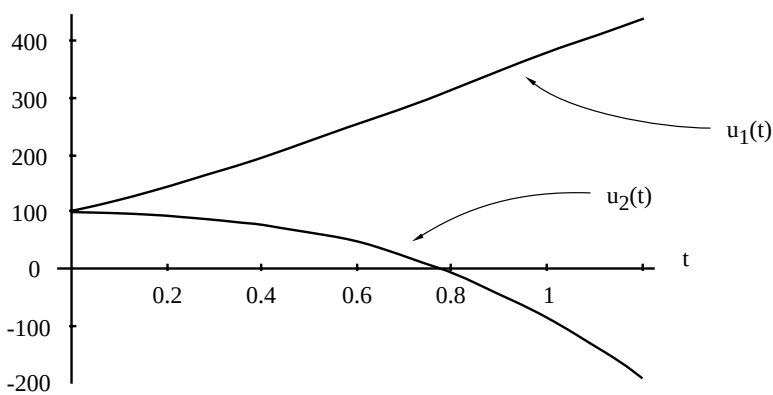


FIGURE 7.4.2

Therefore, the prey species will become extinct at the value of t for which $u_2(t) = 0$—i.e., when

$$100e^t(\cos t - \sin t) = 0 \implies \cos t = \sin t \implies t = \frac{\pi}{4}.$$

Exercises for section 7.4

7.4.1. Suppose that $\mathbf{A}_{n\times n}$ is diagonalizable, and let $\mathbf{P} = [\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n]$ be a matrix whose columns are a complete set of linearly independent eigenvectors corresponding to eigenvalues λ_i. Show that the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, can be written as

$$\mathbf{u}(t) = \xi_1 e^{\lambda_1 t}\mathbf{x}_1 + \xi_2 e^{\lambda_2 t}\mathbf{x}_2 + \cdots + \xi_n e^{\lambda_n t}\mathbf{x}_n$$

in which the coefficients ξ_i satisfy the algebraic system $\mathbf{P}\boldsymbol{\xi} = \mathbf{c}$.

7.4.2. Using only the eigenvalues, determine the long-run behavior of the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$ for each of the following matrices.

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

7.4.3. **Competing Species.** Consider two species that coexist in the same environment but compete for the same resources. Suppose that the population of each species increases proportionally to the number of its own kind but decreases proportionally to the number in the competing species—say that the population of each species increases at a rate equal to twice its existing number but decreases at a rate equal to the number in the other population. Suppose that there are initially 100 of species I and 200 of species II.

(a) Determine the number of each species at all future times.
(b) Determine which species is destined to become extinct, and compute the time to extinction.

7.4.4. **Cooperating Species.** Consider two species that survive in a symbiotic relationship in the sense that the population of each species decreases at a rate equal to its existing number but increases at a rate equal to the existing number in the other population.

(a) If there are initially 200 of species I and 400 of species II, determine the number of each species at all future times.
(b) Discuss the long-run behavior of each species.

7.5 NORMAL MATRICES

A matrix $\mathbf{A}$ is diagonalizable if and only if $\mathbf{A}$ possesses a complete independent set of eigenvectors, and if such a complete set is used for columns of $\mathbf{P}$, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$ is diagonal (p. 507). But even when $\mathbf{A}$ possesses a complete independent set of eigenvectors, there's no guarantee that a complete *orthonormal* set of eigenvectors can be found. In other words, there's no assurance that $\mathbf{P}$ can be taken to be unitary (or orthogonal). And the Gram–Schmidt procedure (p. 309) doesn't help—Gram–Schmidt can turn a basis of eigenvectors into an orthonormal basis but not into an orthonormal basis of eigenvectors. So when (or how) are complete orthonormal sets of eigenvectors produced? In other words, when is $\mathbf{A}$ *unitarily* similar to a diagonal matrix?

Unitary Diagonalization

$\mathbf{A} \in \mathcal{C}^{n\times n}$ is unitarily similar to a diagonal matrix (i.e., $\mathbf{A}$ has a complete orthonormal set of eigenvectors) if and only if $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^*$, in which case $\mathbf{A}$ is said to be a ***normal matrix.***

- Whenever $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$ with $\mathbf{U}$ unitary and $\mathbf{D}$ diagonal, the columns of $\mathbf{U}$ must be a complete orthonormal set of eigenvectors for $\mathbf{A}$, and the diagonal entries of $\mathbf{D}$ are the associated eigenvalues.

Proof. If $\mathbf{A}$ is normal with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then $\mathbf{A} - \lambda_k\mathbf{I}$ is also normal. All normal matrices are RPN (range is perpendicular to nullspace, p. 409), so there is a unitary matrix $\mathbf{U}_k$ such that

$$\mathbf{U}_k^*(\mathbf{A} - \lambda_k\mathbf{I})\mathbf{U}_k = \begin{pmatrix} \mathbf{C}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{(by (5.11.15) on p. 408)}$$

or, equivalently

$$\mathbf{U}_k^*\mathbf{A}\mathbf{U}_k = \begin{pmatrix} \mathbf{C}_k + \lambda_k\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda_k\mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{k-1} & \mathbf{0} \\ \mathbf{0} & \lambda_k\mathbf{I} \end{pmatrix},$$

where $\mathbf{C}_k$ is nonsingular and $\mathbf{A}_{k-1} = \mathbf{C}_k + \lambda_k\mathbf{I}$. Note that $\lambda_k \notin \sigma(\mathbf{A}_{k-1})$ (otherwise $\mathbf{A}_{k-1} - \lambda_k\mathbf{I} = \mathbf{C}_k$ would be singular), so $\sigma(\mathbf{A}_{k-1}) = \{\lambda_1, \lambda_2, \ldots, \lambda_{k-1}\}$ (Exercise 7.1.4). Because $\mathbf{A}_{k-1}$ is also normal, the same argument can be repeated with $\mathbf{A}_{k-1}$ and λ_{k-1} in place $\mathbf{A}$ and λ_k to insure the existence of a unitary matrix $\mathbf{U}_{k-1}$ such that

$$\mathbf{U}_{k-1}^*\mathbf{A}_{k-1}\mathbf{U}_{k-1} = \begin{pmatrix} \mathbf{A}_{k-2} & \mathbf{0} \\ \mathbf{0} & \lambda_{k-1}\mathbf{I} \end{pmatrix},$$

where $\mathbf{A}_{k-2}$ is normal and $\sigma(\mathbf{A}_{k-2}) = \{\lambda_1, \lambda_2, \ldots, \lambda_{k-2}\}$. After k such repetitions, $\mathbf{U}_k \begin{pmatrix} \mathbf{U}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \cdots \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \mathbf{U}$ is a unitary matrix such that

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \begin{pmatrix} \lambda_1\mathbf{I}_{a_1} & 0 & \cdots & 0 \\ 0 & \lambda_2\mathbf{I}_{a_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k\mathbf{I}_{a_k} \end{pmatrix} = \mathbf{D}, \quad a_i = alg\ mult_{\mathbf{A}}(\lambda_i). \tag{7.5.1}$$

Conversely, if there is a unitary matrix $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D}$ is diagonal, then $\mathbf{A}^*\mathbf{A} = \mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^* = \mathbf{U} = \mathbf{U}\mathbf{D}\mathbf{D}^*\mathbf{U}^* = \mathbf{A}\mathbf{A}^*$, so $\mathbf{A}$ is normal. ∎

Caution! While it's true that normal matrices possess a complete orthonormal set of eigenvectors, not all complete independent sets of eigenvectors of a normal $\mathbf{A}$ are orthonormal (or even orthogonal)—see Exercise 7.5.6. Below are some things that are true.

Properties of Normal Matrices

If $\mathbf{A}$ is a normal matrix with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, then

- $\mathbf{A}$ is RPN—i.e., $R(\mathbf{A}) \perp N(\mathbf{A})$ (see p. 408).
- Eigenvectors corresponding to distinct eigenvalues are orthogonal. In other words,

$$N(\mathbf{A} - \lambda_i\mathbf{I}) \perp N(\mathbf{A} - \lambda_j\mathbf{I}) \quad \text{for } \lambda_i \neq \lambda_j. \tag{7.5.2}$$

- The spectral theorems (7.2.7) and (7.3.6) on pp. 517 and 526 hold, but the spectral projectors $\mathbf{G}_i$ on p. 529 specialize to become *orthogonal* projectors because $R(\mathbf{A} - \lambda_i\mathbf{I}) \perp N(\mathbf{A} - \lambda_i\mathbf{I})$ for each λ_i.

Proof of (7.5.2). If $\mathbf{A}$ is normal, so is $\mathbf{A} - \lambda_j\mathbf{I}$, and hence $\mathbf{A} - \lambda_j\mathbf{I}$ is RPN. Consequently, $N(\mathbf{A} - \lambda_j\mathbf{I})^* = N(\mathbf{A} - \lambda_j\mathbf{I})$—recall (5.11.14) from p. 408. If $(\lambda_i, \mathbf{x}_i)$ and $(\lambda_j, \mathbf{x}_j)$ are distinct eigenpairs, then $(\mathbf{A} - \lambda_j\mathbf{I})^*\mathbf{x}_j = \mathbf{0}$, and $0 = \mathbf{x}_j^*(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{x}_i = \mathbf{x}_j^*\mathbf{A}\mathbf{x}_i - \mathbf{x}_j^*\lambda_j\mathbf{x}_i = (\lambda_i - \lambda_j)\mathbf{x}_j^*\mathbf{x}_i$ implies $0 = \mathbf{x}_j^*\mathbf{x}_i$. ∎

Several common types of matrices are normal. For example, real-symmetric and hermitian matrices are normal, real skew-symmetric and skew-hermitian matrices are normal, and orthogonal and unitary matrices are normal. By virtue of being normal, these kinds of matrices inherit all of the above properties, but it's worth looking a bit closer at the real-symmetric and hermitian matrices because they have some special eigenvalue properties.

If $\mathbf{A}$ is real symmetric or hermitian, and if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $\mathbf{x}^*\mathbf{x} \neq 0$, and $\lambda\mathbf{x} = \mathbf{A}\mathbf{x}$ implies $\overline{\lambda}\mathbf{x}^* = \mathbf{x}^*\mathbf{A}^*$, so

$$\mathbf{x}^*\mathbf{x}(\lambda - \overline{\lambda}) = \mathbf{x}^*(\lambda - \overline{\lambda})\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x} - \mathbf{x}^*\mathbf{A}^*\mathbf{x} = 0 \implies \lambda = \overline{\lambda}.$$

In other words, eigenvalues of real-symmetric and hermitian matrices are real. A similar argument (Exercise 7.5.4) shows that the eigenvalues of a real skew-symmetric or skew-hermitian matrix are pure imaginary numbers.

Eigenvectors for a hermitian $\mathbf{A} \in \mathcal{C}^{n\times n}$ may have to involve complex numbers, but a real-symmetric matrix possesses a complete orthonormal set of *real* eigenvectors. Consequently, the real-symmetric case can be distinguished by observing that $\mathbf{A}$ is real symmetric if and only if $\mathbf{A}$ is *orthogonally* similar to a real-diagonal matrix $\mathbf{D}$. Below is a summary of these observations.

Symmetric and Hermitian Matrices

In addition to the properties inherent to all normal matrices,

- Real-symmetric and hermitian matrices have real eigenvalues. (7.5.3)
- $\mathbf{A}$ is real symmetric if and only if $\mathbf{A}$ is *orthogonally* similar to a real-diagonal matrix $\mathbf{D}$—i.e., $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D}$ for some orthogonal $\mathbf{P}$.
- Real skew-symmetric and skew-hermitian matrices have pure imaginary eigenvalues.

Example 7.5.1

Largest and Smallest Eigenvalues. Since the eigenvalues of a hermitian matrix $\mathbf{A}_{n\times n}$ are real, they can be ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

Problem: Explain why the largest and smallest eigenvalues can be described as

$$\lambda_1 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} \quad \text{and} \quad \lambda_n = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x}. \tag{7.5.4}$$

Solution: There is a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ or, equivalently, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$. Since $\|\mathbf{x}\|_2 = 1 \Longleftrightarrow \|\mathbf{y}\|_2 = 1$ for $\mathbf{y} = \mathbf{U}^*\mathbf{x}$,

$$\max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} = \max_{\|\mathbf{y}\|_2=1} \mathbf{y}^*\mathbf{D}\mathbf{y} = \max_{\|\mathbf{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \leq \max_{\|\mathbf{y}\|_2=1} \lambda_1 \sum_{i=1}^{n} |y_i|^2 = \lambda_1$$

with equality being attained when $\mathbf{x}$ is an eigenvector of unit norm associated with λ_1. The expression for the smallest eigenvalue λ_n is obtained by writing

$$\min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x} = \min_{\|\mathbf{y}\|_2=1} \mathbf{y}^*\mathbf{D}\mathbf{y} = \min_{\|\mathbf{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \geq \min_{\|\mathbf{y}\|_2=1} \lambda_n \sum_{i=1}^{n} |y_i|^2 = \lambda_n,$$

where equality is attained at an eigenvector of unit norm associated with λ_n.
Note: The characterizations in (7.5.4) often appear in the equivalent forms

$$\lambda_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}} \quad \text{and} \quad \lambda_n = \min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}.$$

Consequently, $\lambda_1 \geq (\mathbf{x}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x}) \geq \lambda_n$ for all $\mathbf{x} \neq \mathbf{0}$. The term $\mathbf{x}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x}$ is referred to as a ***Rayleigh quotient*** in honor of the famous English physicist John William Strutt (1842–1919) who became Baron Rayleigh in 1873.

It's only natural to wonder if the intermediate eigenvalues of a hermitian matrix have representations similar to those for the extreme eigenvalues as described in (7.5.4). Ernst Fischer (1875–1954) gave the answer for matrices in 1905, and Richard Courant (1888–1972) provided extensions for infinite-dimensional operators in 1920.

Courant–Fischer Theorem

The eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of a hermitian matrix $\mathbf{A}_{n\times n}$ are

$$\lambda_i = \max_{\dim \mathcal{V}=i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x} \quad \text{and} \quad \lambda_i = \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{A}\mathbf{x}. \tag{7.5.5}$$

When $i = 1$ in the min-max formula and when $i = n$ in the max-min formula, $\mathcal{V} = \mathcal{C}^n$, so these cases reduce to the equations in (7.5.4). Alternate max-min and min-max formulas are given in Exercise 7.5.12.

Proof. Only the min-max characterization is proven—the max-min proof is analogous (Exercise 7.5.11). As shown in Example 7.5.1, a change of coordinates $\mathbf{y} = \mathbf{U}^*\mathbf{x}$ with a unitary $\mathbf{U}$ such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ has the effect of replacing $\mathbf{A}$ by $\mathbf{D}$, so we need only establish that

$$\lambda_i = \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{y}\in\mathcal{V} \\ \|\mathbf{y}\|_2=1}} \mathbf{y}^*\mathbf{D}\mathbf{y}.$$

For a subspace $\mathcal{V}$ of dimension $n-i+1$, let $\mathcal{S}_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V},\ \|\mathbf{y}\|_2 = 1\}$, and let

$$\mathcal{S}'_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V} \cap \mathcal{F},\ \|\mathbf{y}\|_2 = 1\}, \quad \text{where} \quad \mathcal{F} = span\,\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\}.$$

Note that $\mathcal{V} \cap \mathcal{F} \neq \mathbf{0}$, for otherwise $\dim(\mathcal{V} + \mathcal{F}) = \dim \mathcal{V} + \dim \mathcal{F} = n + 1$, which is impossible. In other words, $\mathcal{S}'_{\mathcal{V}}$ contains those vectors of $\mathcal{S}_{\mathcal{V}}$ of the form $\mathbf{y} = (y_1, \ldots, y_i, 0, \ldots, 0)^T$ with $\sum_{j=1}^{i} |y_j|^2 = 1$. So for each subspace $\mathcal{V}$ with $\dim \mathcal{V} = n - i + 1$,

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=1}^{i} \lambda_j |y_j|^2 \geq \lambda_i \sum_{j=1}^{i} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}'_{\mathcal{V}}.$$

Since $\mathcal{S}'_{\mathcal{V}} \subseteq \mathcal{S}_{\mathcal{V}}$, it follows that $\max_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \max_{\mathcal{S}'_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i$, and hence

$$\min_{\mathcal{V}} \max_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i.$$

But this inequality is reversible because if $\tilde{\mathcal{V}} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}\}^{\perp}$, then every $\mathbf{y} \in \tilde{\mathcal{V}}$ has the form $\mathbf{y} = (0, \ldots, 0, y_i, \ldots, y_n)^T$, and hence

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=i}^{n} \lambda_j |y_j|^2 \leq \lambda_i \sum_{j=i}^{n} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}_{\tilde{\mathcal{V}}}.$$

So $\min\limits_{\mathcal{V}} \max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \max\limits_{\mathcal{S}_{\tilde{\mathcal{V}}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \lambda_i$, and thus $\min\limits_{\mathcal{V}} \max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} = \lambda_i$. ∎

The value of the Courant–Fischer theorem is its ability to produce inequalities concerning eigenvalues of hermitian matrices without involving the associated eigenvectors. This is illustrated in the following two important examples.

Example 7.5.2

Eigenvalue Perturbations. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of a hermitian $\mathbf{A} \in \mathcal{C}^{n\times n}$, and suppose $\mathbf{A}$ is perturbed by a hermitian $\mathbf{E}$ with eigenvalues $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_n$ to produce $\mathbf{B} = \mathbf{A} + \mathbf{E}$, which is also hermitian.

Problem: If $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ are the eigenvalues of $\mathbf{B}$, explain why

$$\lambda_i + \epsilon_1 \geq \beta_i \geq \lambda_i + \epsilon_n \quad \text{for each } i. \tag{7.5.6}$$

Solution: If $\mathbf{U}$ is a unitary matrix such that $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, then $\tilde{\mathbf{B}} = \mathbf{U}^*\mathbf{B}\mathbf{U}$ and $\tilde{\mathbf{E}} = \mathbf{U}^*\mathbf{E}\mathbf{U}$ have the same eigenvalues as $\mathbf{B}$ and $\mathbf{E}$, respectively, and $\tilde{\mathbf{B}} = \mathbf{D} + \tilde{\mathbf{E}}$. For $\mathbf{x} \in \mathcal{F} = span\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (x_1, \ldots, x_i, 0, \ldots, 0)^T \quad \text{and} \quad \mathbf{x}^*\mathbf{D}\mathbf{x} = \sum_{j=1}^{i} \lambda_j |x_j|^2 \geq \lambda_i \sum_{j=1}^{i} |x_j|^2 = \lambda_i,$$

so applying the max-min part of the Courant–Fischer theorem to $\tilde{\mathbf{B}}$ yields

$$\begin{aligned}\beta_i &= \max_{\dim \mathcal{V}=i} \min_{\substack{\mathbf{x}\in\mathcal{V}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \geq \min_{\substack{\mathbf{x}\in\mathcal{F}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \min_{\substack{\mathbf{x}\in\mathcal{F}\\ \|\mathbf{x}\|_2=1}} \left(\mathbf{x}^*\mathbf{D}\mathbf{x} + \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x}\right)\\ &\geq \min_{\substack{\mathbf{x}\in\mathcal{F}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{D}\mathbf{x} + \min_{\substack{\mathbf{x}\in\mathcal{F}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} \geq \lambda_i + \min_{\substack{\mathbf{x}\in\mathcal{C}^n\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} = \lambda_i + \epsilon_n,\end{aligned}$$

where the last equality is the result of the "min" part of (7.5.4). Similarly, for $\mathbf{x} \in \mathcal{T} = span\{\mathbf{e}_i, \ldots, \mathbf{e}_n\}$ with $\|\mathbf{x}\|_2 = 1$, we have $\mathbf{x}^*\mathbf{D}\mathbf{x} \leq \lambda_i$, and

$$\begin{aligned}\beta_i &= \min_{\dim \mathcal{V}=n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \leq \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \|\mathbf{x}\|_2=1}} \left(\mathbf{x}^*\mathbf{D}\mathbf{x} + \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x}\right)\\ &\leq \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\mathbf{D}\mathbf{x} + \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} \leq \lambda_i + \max_{\substack{\mathbf{x}\in\mathcal{C}^n\\ \|\mathbf{x}\|_2=1}} \mathbf{x}^*\tilde{\mathbf{E}}\mathbf{x} = \lambda_i + \epsilon_1.\end{aligned}$$

Note: Because $\mathbf{E}$ often represents an error, only $\|\mathbf{E}\|$ (or an estimate thereof) is known. But for every matrix norm, $|\epsilon_j| \le \|\mathbf{E}\|$ for each j (Example 7.1.4, p. 497). Since the ϵ_j 's are real, $-\|\mathbf{E}\| \le \epsilon_j \le \|\mathbf{E}\|$, so (7.5.6) guarantees that

$$\lambda_i - \|\mathbf{E}\| \le \beta_i \le \lambda_i + \|\mathbf{E}\|. \tag{7.5.7}$$

In other words,

- the eigenvalues of a hermitian matrix $\mathbf{A}$ are perfectly conditioned because a hermitian perturbation $\mathbf{E}$ changes no eigenvalue of $\mathbf{A}$ by more than $\|\mathbf{E}\|$.

It's interesting to compare (7.5.7) with the Bauer–Fike bound of Example 7.3.2 (p. 528). When $\mathbf{A}$ is hermitian, (7.3.10) reduces to $\min_{\lambda_i \in \sigma(\mathbf{A})} |\beta - \lambda_i| \le \|\mathbf{E}\|$ because $\mathbf{P}$ can be made unitary, so, for induced matrix norms, $\kappa(\mathbf{P}) = 1$. The two results differ in that Bauer–Fike does not assume $\mathbf{E}$ and $\mathbf{B}$ are hermitian.

Example 7.5.3

Interlaced Eigenvalues. For a hermitian matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$ with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$, and for $\mathbf{c} \in \mathcal{C}^{n\times 1}$, let $\mathbf{B}$ be the bordered matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{c} \\ \mathbf{c}^* & \alpha \end{pmatrix}_{n+1\times n+1} \quad \text{with eigenvalues} \quad \beta_1 \ge \beta_2 \ge \cdots \ge \beta_n \ge \beta_{n+1}.$$

Problem: Explain why the eigenvalues of $\mathbf{A}$ interlace with those of $\mathbf{B}$ in that

$$\beta_1 \ge \lambda_1 \ge \beta_2 \ge \lambda_2 \ge \cdots \ge \beta_n \ge \lambda_n \ge \beta_{n+1}. \tag{7.5.8}$$

Solution: To see that $\beta_i \ge \lambda_i \ge \beta_{i+1}$ for $1 \le i \le n$, let $\mathbf{U}$ be a unitary matrix such that $\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Since $\mathbf{V} = \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$ is also unitary, the eigenvalues of $\mathbf{B}$ agree with those of

$$\tilde{\mathbf{B}} = \mathbf{V}^*\mathbf{B}\mathbf{V} = \begin{pmatrix} \mathbf{D} & \mathbf{y} \\ \mathbf{y}^* & \alpha \end{pmatrix}, \quad \text{where} \quad \mathbf{y} = \mathbf{U}^*\mathbf{c}.$$

For $\mathbf{x} \in \mathcal{F} = span\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\} \subset \mathcal{C}^{n+1\times 1}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (x_1, \ldots, x_i, 0, \ldots, 0)^T \quad \text{and} \quad \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \sum_{j=1}^{n} \lambda_j |x_j|^2 \ge \lambda_i \sum_{j=1}^{n} |x_j|^2 = \lambda_i,$$

so applying the max-min part of the Courant–Fisher theorem to $\tilde{\mathbf{B}}$ yields

$$\beta_i = \max_{\dim \mathcal{V} = i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \ge \min_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \ge \lambda_i.$$

For $\mathbf{x} \in \mathcal{T} = span\{\mathbf{e}_{i-1}, \mathbf{e}_i, \ldots, \mathbf{e}_n\} \subset \mathcal{C}^{n+1\times 1}$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathbf{x} = (0, \ldots, 0, x_{i-1}, \ldots, x_n, 0)^T \text{ and } \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} = \sum_{j=i-1}^{n} \lambda_j |x_j|^2 \le \lambda_{i-1} \sum_{j=i}^{n} |x_j|^2 = \lambda_{i-1},$$

so the min-max part of the Courant–Fisher theorem produces

$$\beta_i = \min_{\dim \mathcal{V} = n-i+2} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \le \max_{\substack{\mathbf{x}\in\mathcal{F} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^*\tilde{\mathbf{B}}\mathbf{x} \le \lambda_{i-1}.$$

Note: If $\mathbf{A}$ is any $n \times n$ principal submatrix of $\mathbf{B}$, then (7.5.8) still holds because each principal submatrix can be brought to the upper-left-hand corner by a similarity transformation $\mathbf{P}^T\mathbf{B}\mathbf{P}$, where $\mathbf{P}$ is a permutation matrix. In other words,

- the eigenvalues of an $n+1 \times n+1$ hermitian matrix are interlaced with the eigenvalues of each of its $n \times n$ principal submatrices.

For $\mathbf{A} \in \mathcal{C}^{m\times n}$ (or $\Re^{m\times n}$), the products $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ (or $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$) are hermitian (or real symmetric), so they are diagonalizable by a unitary (or orthogonal) similarity transformation, and their eigenvalues are necessarily real. But in addition to being real, the eigenvalues of these matrices are always nonnegative. For example, if $(\lambda, \mathbf{x})$ is an eigenpair of $\mathbf{A}^*\mathbf{A}$, then $\lambda = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2 \ge 0$, and similarly for the other products. In fact, these λ's are the squares of the singular values for $\mathbf{A}$ developed in §5.12 (p. 411) because if

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{m\times n} \mathbf{V}^*$$

is a singular value decomposition, where $\mathbf{D} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$ contains the nonzero singular values of $\mathbf{A}$, then

$$\mathbf{V}^*\mathbf{A}^*\mathbf{A}\mathbf{V} = \begin{pmatrix} \mathbf{D}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{7.5.9}$$

and this means that $(\sigma_i^2, \mathbf{v}_i)$ for $i = 1, 2, \ldots, r$ is an eigenpair for $\mathbf{A}^*\mathbf{A}$. In other words, *the nonzero singular values of* $\mathbf{A}$ *are precisely the positive square roots of the nonzero eigenvalues of* $\mathbf{A}^*\mathbf{A}$, *and right-hand singular vectors* $\mathbf{v}_i$ *of* $\mathbf{A}$ *are particular eigenvectors of* $\mathbf{A}^*\mathbf{A}$. Note that this establishes the uniqueness of the σ_i's (but not the $\mathbf{v}_i$'s), and pay attention to the fact that the number of zero singular values of $\mathbf{A}$ need not agree with the number of zero eigenvalues of $\mathbf{A}^*\mathbf{A}$—e.g., $\mathbf{A}_{1\times 2} = (1, 1)$ has no zero singular values, but $\mathbf{A}^*\mathbf{A}$ has one zero eigenvalue. The same game can be played with $\mathbf{A}\mathbf{A}^*$ in place of $\mathbf{A}^*\mathbf{A}$ to argue that the nonzero singular values of $\mathbf{A}$ are the positive square roots of

the nonzero eigenvalues of $\mathbf{AA}^*$, and left-hand singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ are particular eigenvectors of $\mathbf{AA}^*$.

Caution! The statement that right-hand singular vectors $\mathbf{v}_i$ of $\mathbf{A}$ are eigenvectors of $\mathbf{A}^*\mathbf{A}$ and left-hand singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ are eigenvectors of $\mathbf{AA}^*$ is a one-way street—it doesn't mean that just any orthonormal sets of eigenvectors for $\mathbf{A}^*\mathbf{A}$ and $\mathbf{AA}^*$ can be used as respective right-hand and left-hand singular vectors for $\mathbf{A}$. The columns $\mathbf{v}_i$ of any unitary matrix $\mathbf{V}$ that diagonalizes $\mathbf{A}^*\mathbf{A}$ as in (7.5.9) can serve as right-hand singular vectors for $\mathbf{A}$, but corresponding left-hand singular vectors $\mathbf{u}_i$ are constrained by the relationships

$$\begin{array}{lll} \mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad i = 1, 2, \ldots, r & \Longrightarrow & \mathbf{u}_i = \dfrac{\mathbf{A}\mathbf{v}_i}{\sigma_i} = \dfrac{\mathbf{A}\mathbf{v}_i}{\|\mathbf{A}\mathbf{v}_i\|_2}, \quad i = 1, 2, \ldots, r, \\ \mathbf{u}_i^*\mathbf{A} = \mathbf{0}, \quad i = r+1, \ldots, m & \Longrightarrow & span\,\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\} = N(\mathbf{A}^*). \end{array}$$

In other words, the first r left-hand singular vectors for $\mathbf{A}$ are uniquely determined by the first r right-hand singular vectors, while the last $m-r$ left-hand singular vectors can be any orthonormal basis for $N(\mathbf{A}^*)$. If $\mathbf{U}$ is constructed from $\mathbf{V}$ as described above, then $\mathbf{U}$ is guaranteed to be unitary because for

$$\mathbf{U} = \left[\mathbf{u}_1 \cdots \mathbf{u}_r | \mathbf{u}_{r+1} \cdots \mathbf{u}_m\right] = \left[\mathbf{U}_1 | \mathbf{U}_2\right] \text{ and } \mathbf{V} = \left[\mathbf{v}_1 \cdots \mathbf{v}_r | \mathbf{v}_{r+1} \cdots \mathbf{v}_n\right] = \left[\mathbf{V}_1 | \mathbf{V}_2\right],$$

$\mathbf{U}_1$ and $\mathbf{U}_2$ each contain orthonormal columns, and, by using (7.5.9),

$$\begin{aligned} R(\mathbf{U}_1) &= R\left(\mathbf{A}\mathbf{V}_1\mathbf{D}^{-1}\right) = R(\mathbf{A}\mathbf{V}_1) = R(\mathbf{A}\mathbf{V}_1\mathbf{D}) = R\left([\mathbf{A}\mathbf{V}_1\mathbf{D}][\mathbf{A}\mathbf{V}_1\mathbf{D}]^*\right) \\ &= R(\mathbf{A}\mathbf{A}^*\mathbf{A}\mathbf{A}^*) = R(\mathbf{A}\mathbf{A}^*) = R(\mathbf{A}) = N(\mathbf{A}^*)^\perp = R(\mathbf{U}_2)^\perp. \end{aligned}$$

The matrix $\mathbf{V}$ is unitary to start with, but, in addition,

$$\begin{aligned} &R(\mathbf{V}_1) = R(\mathbf{V}_1\mathbf{D}) = R([\mathbf{V}_1\mathbf{D}][\mathbf{V}_1\mathbf{D}]^*) = R(\mathbf{A}^*\mathbf{A}) = R(\mathbf{A}^*) \text{ and} \\ &R(\mathbf{V}_2) = R(\mathbf{A}^*)^\perp = N(\mathbf{A}). \end{aligned}$$

These observations are consistent with those established on p. 407 for any URV factorization. Otherwise something would be terribly wrong because an SVD is just a special kind of a URV factorization. Finally, notice that there is nothing special about starting with $\mathbf{V}$ to build a $\mathbf{U}$—we can also take the columns of any unitary $\mathbf{U}$ that diagonalizes $\mathbf{AA}^*$ as left-hand singular vectors for $\mathbf{A}$ and build corresponding right-hand singular vectors in a manner similar to that described above. Below is a summary of the preceding developments concerning singular values together with an additional observation connecting singular values with eigenvalues.

Singular Values and Eigenvalues

For $\mathbf{A} \in \mathcal{C}^{m\times n}$ with $rank\,(\mathbf{A}) = r$, the following statements are valid.

- The nonzero eigenvalues of $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$ are equal and positive.
- The nonzero singular values of $\mathbf{A}$ are the positive square roots of the nonzero eigenvalues of $\mathbf{A}^*\mathbf{A}$ (and $\mathbf{A}\mathbf{A}^*$).
- If $\mathbf{A}$ is normal with nonzero eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$, then the nonzero singular values of $\mathbf{A}$ are $\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_r|\}$.
- Right-hand and left-hand singular vectors for $\mathbf{A}$ are special eigenvectors for $\mathbf{A}^*\mathbf{A}$ and $\mathbf{A}\mathbf{A}^*$, respectively.
- Any complete orthonormal set of eigenvectors $\mathbf{v}_i$ for $\mathbf{A}^*\mathbf{A}$ can serve as a complete set of right-hand singular vectors for $\mathbf{A}$, and a corresponding complete set of left-hand singular vectors is given by $\mathbf{u}_i = \mathbf{A}\mathbf{v}_i / \|\mathbf{A}\mathbf{v}_i\|_2$, $i = 1, 2, \ldots, r$, together with any orthonormal basis $\{\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \ldots, \mathbf{u}_m\}$ for $N(\mathbf{A}^*)$. Similarly, any complete orthonormal set of eigenvectors for $\mathbf{A}\mathbf{A}^*$ can be used as left-hand singular vectors for $\mathbf{A}$, and corresponding right-hand singular vectors can be built in an analogous way.
- The hermitian matrix $\mathbf{B} = \begin{pmatrix} \mathbf{0}_{m\times m} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0}_{n\times n} \end{pmatrix}$ of order $m+n$ has nonzero eigenvalues $\{\pm\sigma_1, \pm\sigma_2, \ldots, \pm\sigma_r\}$ in which $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ are the nonzero singular values of $\mathbf{A}$.

Proof. Only the last point requires proof, and this follows by observing that if λ is an eigenvalue of $\mathbf{B}$, then

$$\begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \implies \left\{ \begin{array}{r} \mathbf{A}\mathbf{x}_2 = \lambda\mathbf{x}_1 \\ \mathbf{A}^*\mathbf{x}_1 = \lambda\mathbf{x}_2 \end{array} \right\} \implies \mathbf{A}^*\mathbf{A}\mathbf{x}_2 = \lambda^2\mathbf{x}_2,$$

so each eigenvalue of $\mathbf{B}$ is the square of a singular value of $\mathbf{A}$. But $\mathbf{B}$ is hermitian with $rank\,(\mathbf{B}) = 2r$, so there are exactly $2r$ nonzero eigenvalues of $\mathbf{B}$. Therefore, each pair $\pm\sigma_i$, $i = 1, 2, \ldots, r$, must be an eigenvalue for $\mathbf{B}$. ∎

Example 7.5.4

Min-Max Singular Values. Since the singular values of $\mathbf{A}$ are the positive square roots of the eigenvalues of $\mathbf{A}^*\mathbf{A}$, and since $\|\mathbf{A}\mathbf{x}\|_2 = (\mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x})^{1/2}$, it's a corollary of the Courant–Fischer theorem (p. 550) that if $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ are the singular values for $\mathbf{A}_{m\times n}$ $(n \leq m)$, then

$$\sigma_i = \max_{\dim \mathcal{V} = i} \min_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \|\mathbf{A}\mathbf{x}\|_2 \quad \text{and} \quad \sigma_i = \min_{\dim \mathcal{V} = n-i+1} \max_{\substack{\mathbf{x}\in\mathcal{V} \\ \|\mathbf{x}\|_2 = 1}} \|\mathbf{A}\mathbf{x}\|_2 .$$

These expressions provide intermediate values between the extremes

$$\sigma_1 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \quad \text{and} \quad \sigma_n = \min_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \qquad \text{(see p. 414).}$$

Exercises for section 7.5

7.5.1. Is $\mathbf{A} = \begin{pmatrix} 5+\mathrm{i} & -2\,\mathrm{i} \\ 2 & 4+2\,\mathrm{i} \end{pmatrix}$ a normal matrix?

7.5.2. Give examples of two distinct classes of normal matrices that are real but not symmetric.

7.5.3. Show that $\mathbf{A} \in \Re^{n\times n}$ is normal and has real eigenvalues if and only if $\mathbf{A}$ is symmetric.

7.5.4. Prove that the eigenvalues of a real skew-symmetric or skew-hermitian matrix must be pure imaginary numbers (i.e., multiples of i).

7.5.5. When trying to decide what's true about matrices and what's not, it helps to think in terms of the following associations.

Hermitian matrices	$\longleftrightarrow$	Real numbers ($z = \overline{z}$).
Skew-hermitian matrices	$\longleftrightarrow$	Imaginary numbers ($z = -\overline{z}$).
Unitary matrices	$\longleftrightarrow$	Points on the unit circle ($z = \mathrm{e}^{\mathrm{i}\theta}$).

For example, the complex function $f(z) = (1-z)(1+z)^{-1}$ maps the imaginary axis in the complex plane to points on the unit circle because $|f(z)|^2 = 1$ whenever $\overline{z} = -z$. It's therefore reasonable to conjecture (as Cayley did in 1846) that if $\mathbf{A}$ is skew hermitian (or real skew symmetric), then

$$f(\mathbf{A}) = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A}) \tag{7.5.10}$$

is unitary (or orthogonal). Prove this is indeed correct. **Note:** Expression (7.5.10) has come to be known as the ***Cayley transformation***.

7.5.6. Show by example that a normal matrix can have a complete independent set of eigenvectors that are not orthonormal, and then explain how every complete independent set of eigenvectors for a normal matrix can be transformed into a complete orthonormal set of eigenvectors.

7.5.7. Construct an example to show that the converse of (7.5.2) is false. In other words, show that it is possible for $N(\mathbf{A} - \lambda_i \mathbf{I}) \perp N(\mathbf{A} - \lambda_j \mathbf{I})$ whenever $i \neq j$ without $\mathbf{A}$ being normal.

7.5.8. Explain why a triangular matrix is normal if and only if it is diagonal.

7.5.9. Use the result of Exercise 7.5.8 to give an alternate proof of the unitary diagonalization theorem given on p. 547.

7.5.10. For a normal matrix $\mathbf{A}$, explain why $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$ if and only if $(\overline{\lambda}, \mathbf{x})$ is an eigenpair for $\mathbf{A}^*$.

7.5.11. To see if you understand the proof of the min-max part of the Courant–Fischer theorem (p. 550), construct an analogous proof for the max-min part of (7.5.5).

7.5.12. The Courant–Fischer theorem has the following alternate formulation.

$$\lambda_i = \max_{\mathbf{v}_1,\ldots,\mathbf{v}_{n-i} \in \mathcal{C}^n} \; \min_{\substack{\mathbf{x} \perp \mathbf{v}_1,\ldots,\mathbf{v}_{n-i} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^* \mathbf{A} \mathbf{x} \quad \text{and} \quad \lambda_i = \min_{\mathbf{v}_1,\ldots,\mathbf{v}_{i-1} \in \mathcal{C}^n} \; \max_{\substack{\mathbf{x} \perp \mathbf{v}_1,\ldots,\mathbf{v}_{i-1} \\ \|\mathbf{x}\|_2 = 1}} \mathbf{x}^* \mathbf{A} \mathbf{x}$$

for $1 < i < n$. To see if you *really* understand the proof of the min-max part of (7.5.5), adapt it to prove the alternate min-max formulation given above.

7.5.13. (a) Explain why every unitary matrix is unitarily similar to a diagonal matrix of the form

$$\mathbf{D} = \begin{pmatrix} e^{i\theta_1} & 0 & \cdots & 0 \\ 0 & e^{i\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{i\theta_n} \end{pmatrix}.$$

(b) Prove that every orthogonal matrix is orthogonally similar to a real block-diagonal matrix of the form

$$\mathbf{B} = \begin{pmatrix} \pm 1 & & & & & & \\ & \ddots & & & & & \\ & & \pm 1 & & & & \\ & & & \cos\theta_1 & \sin\theta_1 & & \\ & & & -\sin\theta_1 & \cos\theta_1 & & \\ & & & & & \ddots & \\ & & & & & \cos\theta_t & \sin\theta_t \\ & & & & & -\sin\theta_t & \cos\theta_t \end{pmatrix}.$$

7.6 POSITIVE DEFINITE MATRICES

Since the symmetric structure of a matrix forces its eigenvalues to be real, what additional property will force all eigenvalues to be *positive* (or perhaps just nonnegative)? To answer this, let's deal with real-symmetric matrices—the hermitian case follows along the same lines. If $\mathbf{A} \in \Re^{n\times n}$ is symmetric, then, as observed above, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{A} = \mathbf{PDP}^T$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is real. If $\lambda_i \geq 0$ for each i, then $\mathbf{D}^{1/2}$ exists, so

$$\mathbf{A} = \mathbf{PDP}^T = \mathbf{PD}^{1/2}\mathbf{D}^{1/2}\mathbf{P}^T = \mathbf{B}^T\mathbf{B} \quad \text{for} \quad \mathbf{B} = \mathbf{D}^{1/2}\mathbf{P}^T,$$

and $\lambda_i > 0$ for each i if and only if $\mathbf{B}$ is nonsingular. Conversely, if $\mathbf{A}$ can be factored as $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, then all eigenvalues of $\mathbf{A}$ are nonnegative because for any eigenpair $(\lambda, \mathbf{x})$,

$$\lambda = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\|\mathbf{B}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \geq 0.$$

Moreover, if $\mathbf{B}$ is nonsingular, then $N(\mathbf{B}) = \mathbf{0} \Longrightarrow \mathbf{B}\mathbf{x} \neq \mathbf{0}$, so $\lambda > 0$. In other words, *a real-symmetric matrix* $\mathbf{A}$ *has nonnegative eigenvalues if and only if* $\mathbf{A}$ *can be factored as* $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, *and all eigenvalues are positive if and only if* $\mathbf{B}$ *is nonsingular*. A symmetric matrix $\mathbf{A}$ whose eigenvalues are positive is called ***positive definite***, and when the eigenvalues are just nonnegative, $\mathbf{A}$ is said to be ***positive semidefinite***.

The use of this terminology is consistent with that introduced in Example 3.10.7 (p. 154), where the term "positive definite" was used to designate symmetric matrices possessing an LU factorization with positive pivots. It was demonstrated in Example 3.10.7 that possessing positive pivots is equivalent to the existence of a *Cholesky factorization* $\mathbf{A} = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R}$ is upper triangular with positive diagonal entries. By the result of the previous paragraph this means that *all eigenvalues of a symmetric matrix* $\mathbf{A}$ *are positive if and only if* $\mathbf{A}$ *has an LU factorization with positive pivots*.

But the pivots are intimately related to the leading principal minor determinants. Recall from Exercise 6.1.16 (p. 474) that if $\mathbf{A}_k$ is the kth leading principal submatrix of $\mathbf{A}_{n\times n}$, then the kth pivot is given by

$$u_{kk} = \begin{cases} \det(\mathbf{A}_1) = a_{11} & \text{for } k = 1, \\ \det(\mathbf{A}_k)/\det(\mathbf{A}_{k-1}) & \text{for } k = 2, 3, \ldots, n. \end{cases}$$

Consequently, *a symmetric matrix is positive definite if and only if each of its leading principal minors is positive*. However, if each leading principal minor is positive, then *all* principal minors must be positive because if $\mathbf{P}_k$ is any principal submatrix of $\mathbf{A}$, then there is a permutation matrix $\mathbf{Q}$ such that

$\mathbf{P}_k$ is a leading principal submatrix in $\mathbf{C} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{P}_k & \star \\ \star & \star \end{pmatrix}$, and, since $\sigma(\mathbf{A}) = \sigma(\mathbf{C})$, we have, with some obvious shorthand notation,

$$\mathbf{A}\text{'s leading pm's} > 0 \Rightarrow \mathbf{A} \text{ pd} \Rightarrow \mathbf{C} \text{ pd} \Rightarrow \det(\mathbf{P}_k) > 0 \Rightarrow \text{all of } \mathbf{A}\text{'s pm's} > 0.$$

Finally, observe that $\mathbf{A}$ *is positive definite if and only if* $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ *for every nonzero* $\mathbf{x} \in \Re^{n\times 1}$. If $\mathbf{A}$ is positive definite, then $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for a nonsingular $\mathbf{B}$, so $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{B}^T\mathbf{B}\mathbf{x} = \|\mathbf{B}\mathbf{x}\|_2^2 \geq 0$ with equality if and only if $\mathbf{B}\mathbf{x} = \mathbf{0}$ or, equivalently, $\mathbf{x} = \mathbf{0}$. Conversely, if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, then for every eigenpair $(\lambda, \mathbf{x})$ we have $\lambda = (\mathbf{x}^T\mathbf{A}\mathbf{x}/\mathbf{x}^T\mathbf{x}) > 0$.

Below is a formal summary of the results for positive definite matrices.

Positive Definite Matrices

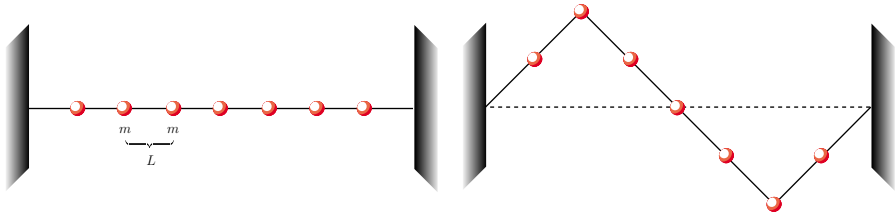
For real-symmetric matrices $\mathbf{A}$, the following statements are equivalent, and any one can serve as the definition of a ***positive definite*** matrix.

- $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for every nonzero $\mathbf{x} \in \Re^{n\times 1}$ (most commonly used as the definition).
- All eigenvalues of $\mathbf{A}$ are positive.
- $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for some nonsingular $\mathbf{B}$.
  - ▷ While $\mathbf{B}$ is not unique, there is one and only one *upper-triangular* matrix $\mathbf{R}$ with positive diagonals such that $\mathbf{A} = \mathbf{R}^T\mathbf{R}$. This is the ***Cholesky factorization*** of $\mathbf{A}$ (Example 3.10.7, p. 154).
- $\mathbf{A}$ has an LU (or LDU) factorization with all pivots being positive.
  - ▷ The LDU factorization is of the form $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T = \mathbf{R}^T\mathbf{R}$, where $\mathbf{R} = \mathbf{D}^{1/2}\mathbf{L}^T$ is the ***Cholesky factor*** of $\mathbf{A}$ (also see p. 154).
- The leading principal minors of $\mathbf{A}$ are positive.
- All principal minors of $\mathbf{A}$ are positive.

For hermitian matrices, replace $(\star)^T$ by $(\star)^*$ and $\Re$ by $\mathcal{C}$.

Example 7.6.1

Vibrating Beads on a String. Consider n small beads, each having mass m, spaced at equal intervals of length L on a very tightly stretched string or wire under a tension T as depicted in Figure 7.6.1. Each bead is initially displaced from its equilibrium position by a small vertical distance—say bead k is displaced by an amount c_k at $t = 0$. The beads are then released so that they can vibrate freely.



Equilibrium Position A Typical Initial Position

FIGURE 7.6.1

Problem: For small vibrations, determine the position of each bead at time $t > 0$ for any given initial configuration.

Solution: The small vibration hypothesis validates the following assumptions.

- The tension T remains constant for all time.
- There is only vertical motion (the horizontal forces cancel each other).
- Only small angles are involved, so the approximation $\sin\theta \approx \tan\theta$ is valid.

Let $y_k(t) = y_k$ be the vertical distance of the kth bead from equilibrium at time t, and set $y_0 = 0 = y_{n+1}$.

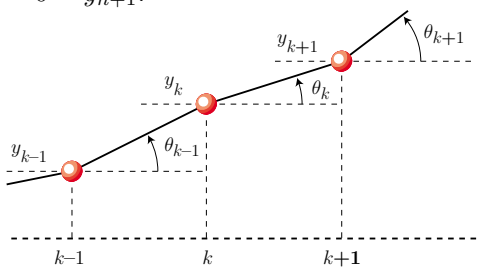


FIGURE 7.6.2

If θ_k is the angle depicted in Figure 7.6.2, the diagram above, then the upward force on the kth bead at time t is $F_u = T\sin\theta_k$, while the downward force is $F_d = T\sin\theta_{k-1}$, so the total force on the kth bead at time t is

$$\begin{aligned} F &= F_u - F_d = T(\sin\theta_k - \sin\theta_{k-1}) \approx T(\tan\theta_k - \tan\theta_{k-1}) \\ &= T\left(\frac{y_{k+1} - y_k}{L} - \frac{y_k - y_{k-1}}{L}\right) = \frac{T}{L}(y_{k-1} - 2y_k + y_{k+1}). \end{aligned}$$

Newton's second law says force = mass × acceleration, so we set

$$my_k'' = \frac{T}{L}(y_{k-1} - 2y_k + y_{k+1}) \implies y_k'' + \frac{T}{mL}(-y_{k-1} + 2y_k - y_{k+1}) = 0 \quad (7.6.1)$$

together with $y_k(0) = c_k$ and $y_k'(0) = 0$ to model the motion of the kth bead. Altogether, equations (7.6.1) represent a system of n second-order linear differential equations, and each is coupled to its neighbors so that no single

equation can be solved in isolation. To extract solutions, the equations must somehow be uncoupled, and here's where matrix diagonalization works its magic. Write equations (7.6.1) in matrix form as

$$\begin{pmatrix} y_1'' \\ y_2'' \\ y_3'' \\ \vdots \\ y_n'' \end{pmatrix} + \frac{T}{mL}\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \text{or} \quad \mathbf{y}'' + \mathbf{A}\mathbf{y} = \mathbf{0}, \tag{7.6.2}$$

with $\mathbf{y}(0) = \mathbf{c} = (c_1 c_2 \cdots c_n)^T$ and $\mathbf{y}'(0) = \mathbf{0}$. Since $\mathbf{A}$ is symmetric, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where the λ_i's are the eigenvalues of $\mathbf{A}$. By making the substitution $\mathbf{y} = \mathbf{P}\mathbf{z}$ (or, equivalently, by changing the coordinate system), (7.6.2) is transformed into

$$\begin{aligned} &\mathbf{z}'' + \mathbf{D}\mathbf{z} = \mathbf{0}, \\ &\mathbf{z}(0) = \mathbf{P}^T\mathbf{c} = \tilde{\mathbf{c}}, \\ &\mathbf{z}'(0) = \mathbf{0}, \end{aligned} \quad \text{or} \quad \begin{pmatrix} z_1'' \\ z_2'' \\ \vdots \\ z_n'' \end{pmatrix} + \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In other words, by changing to a coordinate system defined by a complete set of orthonormal eigenvectors for $\mathbf{A}$, the original system (7.6.2) is completely uncoupled so that each equation $z_k'' + \lambda_k z_k = 0$ with $z_k(0) = \tilde{c}_k$ and $z_k'(0) = 0$ can be solved independently. This helps reveal why diagonalizability is a fundamentally important concept. Recall from elementary differential equations that

$$z_k'' + \lambda_k z_k = 0 \implies z_k(t) = \begin{cases} \alpha_k e^{t\sqrt{-\lambda_k}} + \beta_k e^{-t\sqrt{-\lambda_k}} & \text{when } \lambda_k < 0, \\ \alpha_k \cos\left(t\sqrt{\lambda_k}\right) + \beta_k \sin\left(t\sqrt{\lambda_k}\right) & \text{when } \lambda_k \geq 0. \end{cases}$$

Vibrating beads suggest sinusoidal solutions, so we expect each $\lambda_k > 0$. In other words, the mathematical model would be grossly inconsistent with reality if the symmetric matrix $\mathbf{A}$ in (7.6.2) were not positive definite. It turns out that $\mathbf{A}$ *is* positive definite because there is a Cholesky factorization $\mathbf{A} = \mathbf{R}^T\mathbf{R}$ with

$$\mathbf{R} = \sqrt{\frac{T}{mL}}\begin{pmatrix} r_1 & -1/r_1 & & & \\ & r_2 & -1/r_2 & & \\ & & \ddots & \ddots & \\ & & r_{n-1} & -1/r_{n-1} \\ & & & r_n \end{pmatrix} \quad \text{with} \quad r_k = \sqrt{2 - \frac{k-1}{k}},$$

and thus we are insured that each $\lambda_k > 0$. In fact, since $\mathbf{A}$ is a tridiagonal Toeplitz matrix, the results of Example 7.2.5 (p. 514) can be used to show that

$$\lambda_k = \frac{2T}{mL}\left(1 - \cos\frac{k\pi}{n+1}\right) = \frac{4T}{mL}\sin^2\frac{k\pi}{2(n+1)} \qquad \text{(see Exercise 7.2.18).}$$

Therefore,

$$\left\{\begin{array}{lcl} z_k & = & \alpha_k \cos\left(t\sqrt{\lambda_k}\right) + \beta_k \sin\left(t\sqrt{\lambda_k}\right) \\ z_k(0) & = & \tilde{c}_k \\ z_k'(0) & = & 0 \end{array}\right\} \implies z_k = \tilde{c}_k \cos\left(t\sqrt{\lambda_k}\right), \quad (7.6.3)$$

and for $\mathbf{P} = \left[\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n\right]$,

$$\mathbf{y} = \mathbf{P}\mathbf{z} = z_1\mathbf{x}_1 + z_2\mathbf{x}_2 + \cdots + z_n\mathbf{x}_n = \sum_{j=1}^{n} \left(\tilde{c}_j \cos\left(t\sqrt{\lambda_k}\right)\right)\mathbf{x}_j. \quad (7.6.4)$$

This means that every possible mode of vibration is a combination of modes determined by the eigenvectors $\mathbf{x}_j$. To understand this more clearly, suppose that the beads are initially positioned according to the components of $\mathbf{x}_j$ —i.e., $\mathbf{c} = \mathbf{y}(0) = \mathbf{x}_j$. Then $\tilde{\mathbf{c}} = \mathbf{P}^T\mathbf{c} = \mathbf{P}^T\mathbf{x}_j = \mathbf{e}_j$, so (7.6.3) and (7.6.4) reduce to

$$z_k = \begin{cases} \cos\left(t\sqrt{\lambda_k}\right) & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases} \implies \mathbf{y} = \left(\cos\left(t\sqrt{\lambda_k}\right)\right)\mathbf{x}_j. \quad (7.6.5)$$

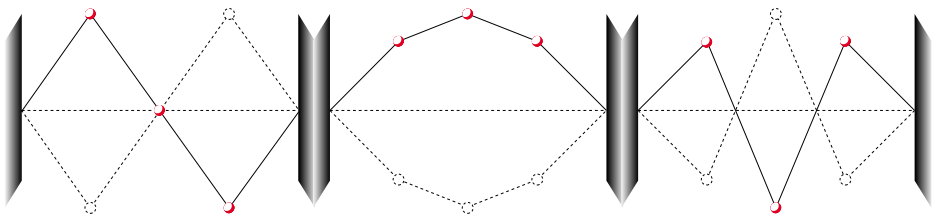
In other words, when $\mathbf{y}(0) = \mathbf{x}_j$, the jth eigenpair $(\lambda_j, \mathbf{x}_j)$ completely determines the mode of vibration because the amplitudes are determined by $\mathbf{x}_j$, and each bead vibrates with a common frequency $f = \sqrt{\lambda_j}/2\pi$. This type of motion (7.6.5) is called a ***normal mode of vibration***. In these terms, equation (7.6.4) translates to say that *every possible mode of vibration is a combination of the normal modes*. For example, when $n = 3$, the matrix in (7.6.2) is

$$\mathbf{A} = \frac{T}{mL}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \text{with} \quad \left\{\begin{array}{l} \lambda_1 = (T/mL)(2) \\ \lambda_2 = (T/mL)(2-\sqrt{2}) \\ \lambda_3 = (T/mL)(2+\sqrt{2}) \end{array}\right\},$$

and a complete orthonormal set of eigenvectors is

$$\mathbf{x}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \frac{1}{2}\begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \frac{1}{2}\begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}.$$

The three corresponding normal modes are shown in Figure 7.6.3.



Mode for $(\lambda_1, \mathbf{x}_1)$ Mode for $(\lambda_2, \mathbf{x}_2)$ Mode for $(\lambda_3, \mathbf{x}_3)$

FIGURE 7.6.3

Example 7.6.2

Discrete Laplacian. According to the laws of physics, the temperature at time t at a point (x, y, z) in a solid body is a function $u(x, y, z, t)$ satisfying the *diffusion equation*

$$\frac{\partial u}{\partial t} = K\nabla^2 u, \quad \text{where} \quad \nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

is the ***Laplacian*** of u and K is a constant of thermal diffusivity. At steady state the temperature at each point does not vary with time, so $\partial u/\partial t = 0$ and $u = u(x, y, z)$ satisfy *Laplace's equation* $\nabla^2 u = 0$. Solutions of this equation are often called *harmonic functions*. The nonhomogeneous equation $\nabla^2 u = f$ (*Poisson's equation*) is addressed in Exercise 7.6.9. To keep things simple, let's confine our attention to the following two-dimensional problem.

Problem: For a square plate as shown in Figure 7.6.4(a), explain how to numerically determine the steady-state temperature at interior grid points when the temperature around the boundary is prescribed to be $u(x, y) = g(x, y)$ for a given function g. In other words, explain how to extract a numerical solution to $\nabla^2 u = 0$ in the interior of the square when $u(x, y) = g(x, y)$ on the square's boundary. This is called a *Dirichlet problem.*[76]

Solution: Discretize the problem by overlaying the plate with a square mesh containing n^2 interior points at equally spaced intervals of length h. As illustrated in Figure 7.6.4(b) for $n = 4$, label the grid points using a rowwise ordering scheme—i.e., label them as you would label matrix entries.

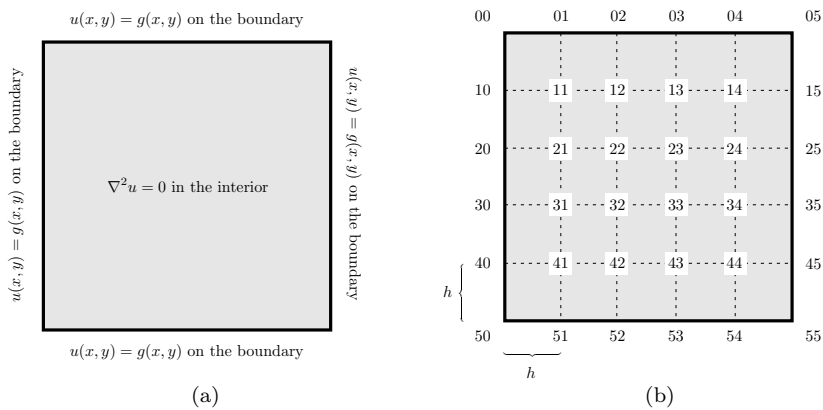


FIGURE 7.6.4

[76] Johann Peter Gustav Lejeune Dirichlet (1805–1859) held the chair at Göttingen previously occupied by Gauss. Because of his work on the convergence of trigonometric series, Dirichlet is generally considered to be the founder of the theory of Fourier series, but much of the groundwork was laid by S. D. Poisson (p. 572) who was Dirichlet's Ph.D. advisor.

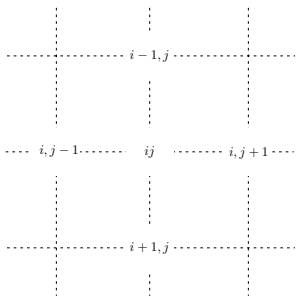
Approximate $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ at the interior grid points (x_i, y_j) by using the second-order centered difference formula (1.4.3) developed on p. 19 to write

$$\begin{aligned}
\frac{\partial^2 u}{\partial x^2}\bigg|_{(x_i,y_j)} &= \frac{u(x_i-h,y_j)-2u(x_i,y_j)+u(x_i+h,y_j)}{h^2}+O(h^2),\\
\frac{\partial^2 u}{\partial y^2}\bigg|_{(x_i,y_j)} &= \frac{u(x_i,y_j-h)-2u(x_i,y_j)+u(x_i,y_j+h)}{h^2}+O(h^2).
\end{aligned} \tag{7.6.6}$$

Adopt the notation $u_{ij} = u(x_i, y_j)$, and add the expressions in (7.6.6) using $\nabla^2 u|_{(x_i,y_j)} = 0$ for interior points (x_i, y_j) to produce

$$4u_{ij} = (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) + O(h^4) \quad \text{for} \quad i,j = 1,2,\ldots,n.$$

In other words, the steady-state temperature at an interior grid point is approximately the average of the steady-state temperatures at the four neighboring grid points as illustrated in Figure 7.6.5.



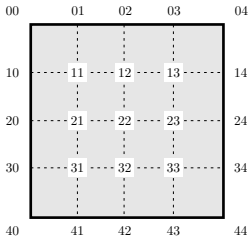
$$u_{ij} = \frac{u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}}{4} + O(h^4)$$

FIGURE 7.6.5

If the $O(h^4)$ terms are neglected, the resulting five-point difference equations,

$$4u_{ij} - (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) = 0 \quad \text{for} \quad i,j = 1,2,\ldots,n,$$

constitute an $n^2 \times n^2$ linear system $\mathbf{Lu} = \mathbf{g}$ in which the unknowns are the u_{ij}'s, and the right-hand side contains boundary values. For example, a mesh with nine interior points produces the 9×9 system in Figure 7.6.6.



$$\left(\begin{array}{ccc|ccc|ccc}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0\\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0\\
\hline
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0\\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0\\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1\\
\hline
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0\\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1\\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{array}\right)
\left(\begin{array}{c}
u_{11}\\ u_{12}\\ u_{13}\\ \hline u_{21}\\ u_{22}\\ u_{23}\\ \hline u_{31}\\ u_{32}\\ u_{33}
\end{array}\right)
=
\left(\begin{array}{c}
g_{01}+g_{10}\\ g_{02}\\ g_{03}+g_{14}\\ \hline g_{20}\\ 0\\ g_{24}\\ \hline g_{30}+g_{41}\\ g_{42}\\ g_{43}+g_{34}
\end{array}\right)$$

FIGURE 7.6.6

The coefficient matrix of this system is the ***discrete Laplacian,*** and in general it has the symmetric block-tridiagonal form

$$\mathbf{L}=\begin{pmatrix} \mathbf{T} & -\mathbf{I} & & & \\ -\mathbf{I} & \mathbf{T} & -\mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I} & \mathbf{T} & -\mathbf{I} \\ & & & -\mathbf{I} & \mathbf{T} \end{pmatrix}_{n^2\times n^2} \quad \text{with} \quad \mathbf{T}=\begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}_{n\times n}.$$

In addition, $\mathbf{L}$ is positive definite. In fact, the discrete Laplacian is a primary example of how positive definite matrices arise in practice. Note that $\mathbf{L}$ is the two-dimensional version of the one-dimensional finite-difference matrix in Example 1.4.1 (p. 19).

Problem: Show $\mathbf{L}$ is positive definite by explicitly exhibiting its eigenvalues.

Solution: Example 7.2.5 (p. 514) insures that the n eigenvalues of $\mathbf{T}$ are

$$\lambda_i = 4 - 2\cos\left(\frac{i\pi}{n+1}\right), \quad i = 1, 2, \ldots, n. \tag{7.6.7}$$

If $\mathbf{U}$ is an orthogonal matrix such that $\mathbf{U}^T\mathbf{T}\mathbf{U} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and if $\mathbf{B}$ is the $n^2 \times n^2$ block-diagonal orthogonal matrix

$$\mathbf{B} = \begin{pmatrix} \mathbf{U} & 0 & \cdots & 0 \\ 0 & \mathbf{U} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{U} \end{pmatrix}, \quad \text{then} \quad \mathbf{B}^T\mathbf{L}\mathbf{B} = \tilde{\mathbf{L}} = \begin{pmatrix} \mathbf{D} & -\mathbf{I} & & & \\ -\mathbf{I} & \mathbf{D} & -\mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I} & \mathbf{D} & -\mathbf{I} \\ & & & -\mathbf{I} & \mathbf{D} \end{pmatrix}.$$

Consider the permutation obtained by placing the numbers $1, 2, \ldots, n^2$ rowwise in a square matrix, and then reordering them by listing the entries columnwise. For example, when $n = 3$ this permutation is generated as follows:

$$\mathbf{v} = (1,2,3,4,5,6,7,8,9) \to \mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \to (1,4,7,2,5,8,3,6,9) = \tilde{\mathbf{v}}.$$

Equivalently, this can be described in terms of wrapping and unwrapping rows by writing $\mathbf{v} \xrightarrow{wrap} \mathbf{A} \longrightarrow \mathbf{A}^T \xrightarrow{unwrap} \tilde{\mathbf{v}}$. If $\mathbf{P}$ is the associated $n^2 \times n^2$ permutation matrix, then

$$\mathbf{P}^T\tilde{\mathbf{L}}\mathbf{P} = \begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{T}_n \end{pmatrix} \quad \text{with} \quad \mathbf{T}_i = \begin{pmatrix} \lambda_i & -1 & & & \\ -1 & \lambda_i & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & \lambda_i & -1 \\ & & & -1 & \lambda_i \end{pmatrix}_{n\times n}.$$

If you try it on the 9×9 case, you will see why it works. Now, $\mathbf{T}_i$ is another tridiagonal Toeplitz matrix, so Example 7.2.5 (p. 514) again applies to yield $\sigma(\mathbf{T}_i) = \{\lambda_i - 2\cos(j\pi/n+1),\ j = 1, 2, \ldots, n\}$. This together with (7.6.7) produces the n^2 eigenvalues of $\mathbf{L}$ as

$$\lambda_{ij} = 4 - 2\left[\cos\left(\frac{i\pi}{n+1}\right) + \cos\left(\frac{j\pi}{n+1}\right)\right], \qquad i, j = 1, 2, \ldots, n,$$

or, by using the identity $1 - \cos\theta = 2\sin^2(\theta/2)$,

$$\lambda_{ij} = 4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right) + \sin^2\left(\frac{j\pi}{2(n+1)}\right)\right], \qquad i, j = 1, 2, \ldots, n. \tag{7.6.8}$$

Since each λ_{ij} is positive, $\mathbf{L}$ must be positive definite. As a corollary, $\mathbf{L}$ is nonsingular, and hence $\mathbf{Lu} = \mathbf{g}$ yields a unique solution for the steady-state temperatures on the square plate (otherwise something would be amiss).

At first glance it's tempting to think that statements about positive definite matrices translate to positive semidefinite matrices simply by replacing the word "positive" by "nonnegative," but this is not always true. When $\mathbf{A}$ has zero eigenvalues (i.e., when $\mathbf{A}$ is singular) there is no LU factorization, and, unlike the positive definite case, having nonnegative leading principal minors doesn't insure that $\mathbf{A}$ is positive semidefinite—e.g., consider $\mathbf{A} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$. The positive definite properties that have semidefinite analogues are listed below.

Positive Semidefinite Matrices

For real-symmetric matrices such that $rank\,(\mathbf{A}_{n\times n}) = r$, the following statements are equivalent, so any one of them can serve as the definition of a ***positive semidefinite*** matrix.

- $\mathbf{x}^T\mathbf{Ax} \geq 0$ for all $\mathbf{x} \in \Re^{n\times 1}$ (the most common definition). (7.6.9)
- All eigenvalues of $\mathbf{A}$ are nonnegative. (7.6.10)
- $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ for some $\mathbf{B}$ with $rank\,(\mathbf{B}) = r$. (7.6.11)
- All principal minors of $\mathbf{A}$ are nonnegative. (7.6.12)

For hermitian matrices, replace $(\star)^T$ by $(\star)^*$ and $\Re$ by $\mathcal{C}$.

Proof of (7.6.9) $\Longrightarrow$ (7.6.10). The hypothesis insures $\mathbf{x}^T\mathbf{Ax} \geq 0$ for eigenvectors of $\mathbf{A}$. If $(\lambda, \mathbf{x})$ is an eigenpair, then $\lambda = \mathbf{x}^T\mathbf{Ax}/\mathbf{x}^T\mathbf{x} = \|\mathbf{Bx}\|_2^2 / \|\mathbf{x}\|_2^2 \geq 0$.

Proof of (7.6.10) $\Longrightarrow$ (7.6.11). Similar to the positive definite case, if each $\lambda_i \geq 0$, write $\mathbf{A} = \mathbf{PD}^{1/2}\mathbf{D}^{1/2}\mathbf{P}^T = \mathbf{B}^T\mathbf{B}$, where $\mathbf{B} = \mathbf{D}^{1/2}\mathbf{P}^T$ has rank r.

Proof of (7.6.11) $\Longrightarrow$ (7.6.12). If $\mathbf{P}_k$ is a principal submatrix of $\mathbf{A}$, then

$$\begin{pmatrix} \mathbf{P}_k & \star \\ \star & \star \end{pmatrix} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{Q}^T\mathbf{B}^T\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{F}^T \\ \star \end{pmatrix} \left[\mathbf{F} \mid \star\right] \quad \Longrightarrow \quad \mathbf{P}_k = \mathbf{F}^T\mathbf{F}$$

for a permutation matrix $\mathbf{Q}$. Thus $\det(\mathbf{P}_k) = \det\left(\mathbf{F}^T\mathbf{F}\right) \geq 0$ (Exercise 6.1.10).

Proof of (7.6.12) $\Longrightarrow$ (7.6.9). If $\mathbf{A}_k$ is the leading $k \times k$ principal submatrix of $\mathbf{A}$, and if $\{\mu_1, \mu_2, \ldots, \mu_k\}$ are the eigenvalues (including repetitions) of $\mathbf{A}_k$, then $\epsilon\mathbf{I} + \mathbf{A}_k$ has eigenvalues $\{\epsilon + \mu_1, \epsilon + \mu_2, \ldots, \epsilon + \mu_k\}$, so, for every $\epsilon > 0$,

$$\det(\epsilon\mathbf{I} + \mathbf{A}_k) = (\epsilon + \mu_1)(\epsilon + \mu_2)\cdots(\epsilon + \mu_k) = \epsilon^k + s_1\epsilon^{k-1} + \cdots + \epsilon s_{k-1} + s_k > 0$$

because s_j is the jth symmetric function of the μ_i's (p. 494), and, by (7.1.6), s_j is the sum of the $j \times j$ principal minors of $\mathbf{A}_k$, which are principal minors of $\mathbf{A}$. In other words, each leading principal minor of $\epsilon\mathbf{I} + \mathbf{A}$ is positive, so $\epsilon\mathbf{I}+\mathbf{A}$ is positive definite by the results on p. 559. Consequently, for each nonzero $\mathbf{x} \in \Re^{n\times 1}$, we must have $\mathbf{x}^T(\epsilon\mathbf{I} + \mathbf{A})\mathbf{x} > 0$ for every $\epsilon > 0$. Let $\epsilon \to 0^+$ (i.e., through positive values) to conclude that $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for each $\mathbf{x} \in \Re^{n\times 1}$. ∎

Quadratic Forms

For a vector $\mathbf{x} \in \Re^{n\times 1}$ and a matrix $\mathbf{A} \in \Re^{n\times n}$, the scalar function defined by

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j \tag{7.6.13}$$

is called a ***quadratic form***. A quadratic form is said to be *positive definite* whenever $\mathbf{A}$ is a positive definite matrix. In other words, (7.6.13) is a positive definite form if and only if $f(\mathbf{x}) > 0$ for all $\mathbf{0} \neq \mathbf{x} \in \Re^{n\times 1}$.

Because $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\left[(\mathbf{A}+\mathbf{A}^T)/2\right]\mathbf{x}$, and because $(\mathbf{A}+\mathbf{A}^T)/2$ is symmetric, the matrix of a quadratic form can always be forced to be symmetric. For this reason it is assumed that the matrix of *every* quadratic form is symmetric. When $\mathbf{x} \in \mathcal{C}^{n\times 1}$, $\mathbf{A} \in \mathcal{C}^{n\times n}$, and $\mathbf{A}$ is hermitian, the expression $\mathbf{x}^H\mathbf{A}\mathbf{x}$ is known as a *complex quadratic form*.

Example 7.6.3

Diagonalization of a Quadratic Form. A quadratic form $f(\mathbf{x}) = \mathbf{x}^T\mathbf{D}\mathbf{x}$ is said to be a *diagonal form* whenever $\mathbf{D}_{n\times n}$ is a diagonal matrix, in which case $\mathbf{x}^T\mathbf{D}\mathbf{x} = \sum_{i=1}^{n} d_{ii}x_i^2$ (there are no cross-product terms). Every quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ can be diagonalized by making a change of variables (coordinates)

$\mathbf{y} = \mathbf{Q}^T\mathbf{x}$. This follows because $\mathbf{A}$ is symmetric, so there is an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, where $\lambda_i \in \sigma(\mathbf{A})$, and setting $\mathbf{y} = \mathbf{Q}^T\mathbf{x}$ (or, equivalently, $\mathbf{x} = \mathbf{Q}\mathbf{y}$) gives

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{Q}^T\mathbf{A}\mathbf{Q}\mathbf{y} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2. \tag{7.6.14}$$

This shows that the nature of the quadratic form is determined by the eigenvalues of $\mathbf{A}$ (which are necessarily real). The effect of diagonalizing a quadratic form in this way is to rotate the standard coordinate system so that in the new coordinate system the graph of $\mathbf{x}^T\mathbf{A}\mathbf{x} = \alpha$ is in "standard form." If $\mathbf{A}$ is positive definite, then all of its eigenvalues are positive (p. 559), so (7.6.14) makes it clear that the graph of $\mathbf{x}^T\mathbf{A}\mathbf{x} = \alpha$ for a constant $\alpha > 0$ is an ellipsoid centered at the origin. Go back and look at Figure 7.2.1 (p. 505), and see Exercise 7.6.4 (p. 571).

Example 7.6.4

Congruence. It's not necessary to solve an eigenvalue problem to diagonalize a quadratic form because a ***congruence transformation*** $\mathbf{C}^T\mathbf{A}\mathbf{C}$ in which $\mathbf{C}$ is nonsingular (but not necessarily orthogonal) can be found that will do the job. A particularly convenient congruence transformation is produced by the LDU factorization for $\mathbf{A}$, which is $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ because $\mathbf{A}$ is symmetric—see Exercise 3.10.9 (p. 157). This factorization is relatively cheap, and the diagonal entries in $\mathbf{D} = \operatorname{diag}(p_1, p_2, \ldots, p_n)$ are the pivots that emerge during Gaussian elimination (p. 154). Setting $\mathbf{y} = \mathbf{L}^T\mathbf{x}$ (or, equivalently, $\mathbf{x} = (\mathbf{L}^T)^{-1}\mathbf{y}$) yields

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} p_i y_i^2.$$

The ***inertia*** of a real-symmetric matrix $\mathbf{A}$ is defined to be the triple (ρ, ν, ζ) in which ρ, ν, and ζ are the respective number of positive, negative, and zero eigenvalues, counting algebraic multiplicities. In 1852 J. J. Sylvester (p. 80) discovered that the inertia of $\mathbf{A}$ is invariant under congruence transformations.

Sylvester's Law of Inertia

Let $\mathbf{A} \cong \mathbf{B}$ denote the fact that real-symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ are congruent (i.e., $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathbf{B}$ for some nonsingular $\mathbf{C}$). Sylvester's law of inertia states that:

$$\mathbf{A} \cong \mathbf{B} \quad \text{if and only if } \mathbf{A} \text{ and } \mathbf{B} \text{ have the same inertia.}$$

Proof.[77] Observe that if $\mathbf{A}_{n\times n}$ is real and symmetric with inertia (p,j,s), then

$$\mathbf{A} \cong \begin{pmatrix} \mathbf{I}_{p\times p} & & \\ & -\mathbf{I}_{j\times j} & \\ & & \mathbf{0}_{s\times s} \end{pmatrix} = \mathbf{E}, \tag{7.6.15}$$

because if $\{\lambda_1,\ldots,\lambda_p,-\lambda_{p+1},\ldots,-\lambda_{p+j},0,\ldots,0\}$ are the eigenvalues of $\mathbf{A}$ (counting multiplicities) with each $\lambda_i > 0$, there is an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \operatorname{diag}(\lambda_1,\ldots,\lambda_p,-\lambda_{p+1},\ldots,-\lambda_{p+j},0,\ldots,0)$, so $\mathbf{C} = \mathbf{P}\mathbf{D}$, where $\mathbf{D} = \operatorname{diag}\left(\lambda_1^{-1/2},\ldots,\lambda_{p+j}^{-1/2},1,\ldots,1\right)$, is nonsingular and $\mathbf{C}^T\mathbf{A}\mathbf{C} = \mathbf{E}$. Let $\mathbf{B}$ be a real-symmetric matrix with inertia (q,k,t) so that

$$\mathbf{B} \cong \begin{pmatrix} \mathbf{I}_{q\times q} & & \\ & -\mathbf{I}_{k\times k} & \\ & & \mathbf{0}_{t\times t} \end{pmatrix} = \mathbf{F}.$$

If $\mathbf{B} \cong \mathbf{A}$, then $\mathbf{F} \cong \mathbf{E}$ (congruence is transitive), so $rank\,(\mathbf{F}) = rank\,(\mathbf{E})$, and hence $s = t$. To show that $p = q$, assume to the contrary that $p > q$, and write $\mathbf{F} = \mathbf{K}^T\mathbf{E}\mathbf{K}$ for some nonsingular $\mathbf{K} = \left(\mathbf{X}_{n\times q}\,|\,\mathbf{Y}_{n\times n-q}\right)$. If $\mathcal{M} = R\,(\mathbf{Y}) \subseteq \Re^n$ and $\mathcal{N} = span\,\{\mathbf{e}_1,\ldots,\mathbf{e}_p\} \subseteq \Re^n$, then using the formula (4.4.19) for the dimension of a sum (p. 205) yields

$$\dim(\mathcal{M}\cap\mathcal{N}) = \dim\mathcal{M} + \dim\mathcal{N} - \dim(\mathcal{M}+\mathcal{N}) = (n-q)+p-\dim(\mathcal{M}+\mathcal{N}) > 0.$$

Consequently, there exists a nonzero vector $\mathbf{x} \in \mathcal{M}\cap\mathcal{N}$. For such a vector,

$$\mathbf{x} \in \mathcal{M} \implies \mathbf{x} = \mathbf{Y}\mathbf{y} = \mathbf{K}\begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix} \implies \mathbf{x}^T\mathbf{E}\mathbf{x} = \left(\mathbf{0}^T\,|\,\mathbf{y}^T\right)\mathbf{F}\begin{pmatrix}\mathbf{0}\\ \mathbf{y}\end{pmatrix} \le 0,$$

and

$$\mathbf{x} \in \mathcal{N} \implies \mathbf{x} = (x_1,\ldots,x_p,0,\ldots,0)^T \implies \mathbf{x}^T\mathbf{E}\mathbf{x} > 0,$$

which is impossible. Therefore, we can't have $p > q$. A similar argument shows that it's also impossible to have $p < q$, so $p = q$. Thus it is proved that if $\mathbf{A} \cong \mathbf{B}$, then $\mathbf{A}$ and $\mathbf{B}$ have the same inertia. Conversely, if $\mathbf{A}$ and $\mathbf{B}$ have inertia (p,j,s), then the argument that produced (7.6.15) yields $\mathbf{A} \cong \mathbf{E} \cong \mathbf{B}$. ∎

[77] The fact that inertia is invariant under congruence is also a corollary of a deeper theorem stating that the eigenvalues of $\mathbf{A}$ vary continuously with the entries. The argument is as follows. Assume $\mathbf{A}$ is nonsingular (otherwise consider $\mathbf{A}+\epsilon\mathbf{I}$ for small ϵ), and set $\mathbf{X}(t) = t\mathbf{Q} + (1-t)\mathbf{Q}\mathbf{R}$ for $t \in [0,1]$, where $\mathbf{C} = \mathbf{Q}\mathbf{R}$ is the QR factorization. Both $\mathbf{X}(t)$ and $\mathbf{Y}(t) = \mathbf{X}^T(t)\mathbf{A}\mathbf{X}(t)$ are nonsingular on $[0,1]$, so continuity of eigenvalues insures that no eigenvalue $\mathbf{Y}(t)$ can cross the origin as t goes from 0 to 1. Hence $\mathbf{Y}(0) = \mathbf{C}^T\mathbf{A}\mathbf{C}$ has the same number of positive (and negative) eigenvalues as $\mathbf{Y}(1) = \mathbf{Q}^T\mathbf{A}\mathbf{Q}$, which is similar to $\mathbf{A}$. Thus $\mathbf{C}^T\mathbf{A}\mathbf{C}$ and $\mathbf{A}$ have the same inertia.

Example 7.6.5

Taylor's theorem in $\Re^n$ says that if f is a smooth real-valued function defined on $\Re^n$, and if $\mathbf{x}_0 \in \Re^{n\times 1}$, then the value of f at $\mathbf{x} \in \Re^{n\times 1}$ is given by

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T\mathbf{g}(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T\mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + O(\|\mathbf{x}-\mathbf{x}_0\|^3),$$

where $\mathbf{g}(\mathbf{x}_0) = \nabla f(\mathbf{x}_0)$ (the gradient of f evaluated at $\mathbf{x}_0$) has components $g_i = \partial f/\partial x_i\Big|_{\mathbf{x}_0}$, and where $\mathbf{H}(\mathbf{x}_0)$ is the ***Hessian matrix*** whose entries are given by $h_{ij} = \partial^2 f/\partial x_i \partial x_j\Big|_{\mathbf{x}_0}$. Just as in the case of one variable, the vector $\mathbf{x}_0$ is called a *critical point* when $\mathbf{g}(\mathbf{x}_0) = \mathbf{0}$. If $\mathbf{x}_0$ is a critical point, then Taylor's theorem shows that $(\mathbf{x}-\mathbf{x}_0)^T\mathbf{H}(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)$ governs the behavior of f at points $\mathbf{x}$ near to $\mathbf{x}_0$. This observation yields the following conclusions regarding local maxima or minima.

- If $\mathbf{x}_0$ is a critical point such that $\mathbf{H}(\mathbf{x}_0)$ is positive definite, then f has a local minimum at $\mathbf{x}_0$.
- If $\mathbf{x}_0$ is a critical point such that $\mathbf{H}(\mathbf{x}_0)$ is *negative definite* (i.e., $\mathbf{z}^T\mathbf{H}\mathbf{z} < 0$ for all $\mathbf{z} \neq \mathbf{0}$ or, equivalently, $-\mathbf{H}$ is positive definite), then f has a local maximum at $\mathbf{x}_0$.

Exercises for section 7.6

7.6.1. Which of the following matrices are positive definite?

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 5 & 1 \\ -1 & 1 & 5 \end{pmatrix}. \quad \mathbf{B} = \begin{pmatrix} 20 & 6 & 8 \\ 6 & 3 & 0 \\ 8 & 0 & 8 \end{pmatrix}. \quad \mathbf{C} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 6 & 2 \\ 2 & 2 & 4 \end{pmatrix}.$$

7.6.2. Spring-Mass Vibrations. Two masses m_1 and m_2 are suspended between three identical springs (with spring constant k) as shown in Figure 7.6.7. Each mass is initially displaced from its equilibrium position by a horizontal distance and released to vibrate freely (assume there is no vertical displacement).

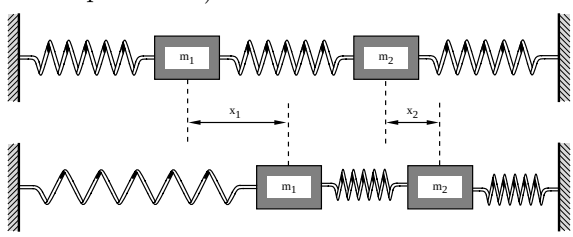


FIGURE 7.6.7

(a) If $x_i(t)$ denotes the horizontal displacement of m_i from equilibrium at time t, show that $\mathbf{M}\mathbf{x}'' = \mathbf{K}\mathbf{x}$, where

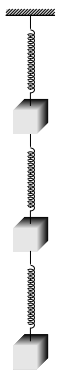
$$\mathbf{M} = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \quad \text{and} \quad \mathbf{K} = k\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(Consider a force directed to the left to be positive.) Notice that the *mass-stiffness equation* $\mathbf{M}\mathbf{x}'' = \mathbf{K}\mathbf{x}$ is the matrix version of Hooke's law $F = kx$, and $\mathbf{K}$ is positive definite.

(b) Look for a solution of the form $\mathbf{x} = e^{i\theta t}\mathbf{v}$ for a constant vector $\mathbf{v}$, and show that this reduces the problem to solving an algebraic equation of the form $\mathbf{K}\mathbf{v} = \lambda\mathbf{M}\mathbf{v}$ (for $\lambda = -\theta^2$). This is called a ***generalized eigenvalue problem*** because when $\mathbf{M} = \mathbf{I}$ we are back to the ordinary eigenvalue problem. The *generalized eigenvalues* λ_1 and λ_2 are the roots of the equation $\det(\mathbf{K} - \lambda\mathbf{M}) = 0$—find them when $k = 1$, $m_1 = 1$, and $m_2 = 2$, and describe the two modes of vibration.

(c) Take $m_1 = m_2 = m$, and apply the technique used in the vibrating beads problem in Example 7.6.1 (p. 559) to determine the normal modes. Compare the results with those of part (b).

7.6.3. Three masses m_1, m_2, and m_3 are suspended on three identical springs (with spring constant k) as shown below. Each mass is initially displaced from its equilibrium position by a vertical distance and then released to vibrate freely.



(a) If $y_i(t)$ denotes the displacement of m_i from equilibrium at time t, show that the mass-stiffness equation is $\mathbf{M}\mathbf{y}'' = \mathbf{K}\mathbf{y}$, where

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix}, \quad \mathbf{K} = k\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}$$

($k_{33} = 1$ is not a mistake!).

(b) Show that $\mathbf{K}$ is positive definite.

(c) Find the normal modes when $m_1 = m_2 = m_3 = m$.

7.6.4. By diagonalizing the quadratic form $13x^2 + 10xy + 13y^2$, show that the rotated graph of $13x^2 + 10xy + 13y^2 = 72$ is an ellipse in standard form as shown in Figure 7.2.1 on p. 505.

7.6.5. Suppose that $\mathbf{A}$ is a real-symmetric matrix. Explain why the signs of the pivots in the LDU factorization for $\mathbf{A}$ reveal the inertia of $\mathbf{A}$.

7.6.6. Consider the quadratic form

$$f(\mathbf{x}) = \frac{1}{9}(-2x_1^2 + 7x_2^2 + 4x_3^2 + 4x_1x_2 + 16x_1x_3 + 20x_2x_3).$$

(a) Find a symmetric matrix $\mathbf{A}$ so that $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$.
(b) Diagonalize the quadratic form using the $\mathbf{LDL}^T$ factorization as described in Example 7.6.4, and determine the inertia of $\mathbf{A}$.
(c) Is this a positive definite form?
(d) Verify the inertia obtained above is correct by computing the eigenvalues of $\mathbf{A}$.
(e) Verify Sylvester's law of inertia by making up a congruence transformation $\mathbf{C}$ and then computing the inertia of $\mathbf{C}^T\mathbf{A}\mathbf{C}$.

7.6.7. Polar Factorization. Explain why each nonsingular $\mathbf{A} \in \mathcal{C}^{n\times n}$ can be uniquely factored as $\mathbf{A} = \mathbf{RU}$, where $\mathbf{R}$ is hermitian positive definite and $\mathbf{U}$ is unitary. This is the matrix analog of the polar form of a complex number $z = re^{i\theta}$, $r > 0$, because 1×1 hermitian positive definite matrices are positive real numbers, and 1×1 unitary matrices are points on the unit circle. **Hint:** First explain why $\mathbf{R} = (\mathbf{AA}^*)^{1/2}$.

7.6.8. Explain why trying to produce better approximations to the solution of the Dirichlet problem in Example 7.6.2 by using finer meshes with more grid points results in an increasingly ill-conditioned linear system $\mathbf{Lu} = \mathbf{g}$.

7.6.9. For a given function f the equation $\nabla^2 u = f$ is called *Poisson's*[78] *equation.* Consider Poisson's equation on a square in two dimensions with Dirichlet boundary conditions. That is,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad \text{with} \quad u(x, y) = g(x, y) \quad \text{on the boundary.}$$

[78] Siméon Denis Poisson (1781–1840) was a prolific French scientist who was originally encouraged to study medicine but was seduced by mathematics. While he was still a teenager, his work attracted the attention of the reigning scientific elite of France such as Legendre, Laplace, and Lagrange. The latter two were originally his teachers (Lagrange was his thesis director) at the École Polytechnique, but they eventually became his friends and collaborators. It is estimated that Poisson published about 400 scientific articles, and his 1811 book *Traité de mécanique* was the standard reference for mechanics for many years. Poisson began his career as an astronomer, but he is primarily remembered for his impact on applied areas such as mechanics, probability, electricity and magnetism, and Fourier series. This seems ironic because he held the chair of "pure mathematics" in the Faculté des Sciences. The next time you find yourself on the streets of Paris, take a stroll on the Rue Denis Poisson, or you can check out Poisson's plaque, along with those of Lagrange, Laplace, and Legendre, on the first stage of the Eiffel Tower.

Discretize the problem by overlaying the square with a regular mesh containing n^2 interior points at equally spaced intervals of length h as explained in Example 7.6.2 (p. 563). Let $f_{ij} = f(x_i, y_j)$, and define $\mathbf{f}$ to be the vector $\mathbf{f} = (f_{11}, f_{12}, \ldots, f_{1n} | f_{21}, f_{22} \ldots, f_{2n} | \cdots | f_{n1}, f_{n2}, \ldots, f_{nn})^T$. Show that the discretization of Poisson's equation produces a system of linear equations of the form $\mathbf{Lu} = \mathbf{g} - h^2\mathbf{f}$, where $\mathbf{L}$ is the discrete Laplacian and where $\mathbf{u}$ and $\mathbf{g}$ are as described in Example 7.6.2.

7.6.10. As defined in Exercise 5.8.15 (p. 380) and discussed in Exercise 7.8.11 (p. 597) the *Kronecker product* (sometimes called *tensor product*, or *direct product*) of matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

Verify that if $\mathbf{I}_n$ is the $n \times n$ identity matrix, and if

$$\mathbf{A}_n = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{n\times n}$$

is the nth-order finite difference matrix of Example 1.4.1 (p. 19), then the discrete Laplacian is given by

$$\mathbf{L}_{n^2\times n^2} = (\mathbf{I}_n \otimes \mathbf{A}_n) + (\mathbf{A}_n \otimes \mathbf{I}_n).$$

Thus we have an elegant matrix connection between the finite difference approximations of the one-dimensional and two-dimensional Laplacians. This formula leads to a simple alternate derivation of (7.6.8)—see Exercise 7.8.12 (p. 598). As you might guess, the discrete three-dimensional Laplacian is

$$\mathbf{L}_{n^3\times n^3} = (\mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{A}_n) + (\mathbf{I}_n \otimes \mathbf{A}_n \otimes \mathbf{I}_n) + (\mathbf{A}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n).$$

7.7 NILPOTENT MATRICES AND JORDAN STRUCTURE

While it's not always possible to diagonalize a matrix $\mathbf{A} \in \mathcal{C}^{m\times m}$ with a similarity transformation, Schur's theorem (p. 508) guarantees that every $\mathbf{A} \in \mathcal{C}^{m\times m}$ is *unitarily* similar to an upper-triangular matrix—say $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$. But other than the fact that the diagonal entries of $\mathbf{T}$ are the eigenvalues of $\mathbf{A}$, there is no pattern to the nonzero part of $\mathbf{T}$. So to what extent can this be remedied by giving up the unitary nature of $\mathbf{U}$? In other words, is there a nonunitary $\mathbf{P}$ for which $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ has a simpler and more predictable pattern than that of $\mathbf{T}$? We have already made the first step in answering this question. The core-nilpotent decomposition (p. 397) says that for every singular matrix $\mathbf{A}$ of index k and rank r, there is a nonsingular matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{L} \end{pmatrix}, \text{ where } rank\,(\mathbf{C}) = r \text{ and } \mathbf{L} \text{ is nilpotent of index } k.$$

Consequently, any further simplification by means of similarity transformations can revolve around $\mathbf{C}$ and $\mathbf{L}$. Let's begin by examining the degree to which nilpotent matrices can be reduced by similarity transformations.

In what follows, let $\mathbf{L}_{n\times n}$ be a nilpotent matrix of index k so that $\mathbf{L}^k = \mathbf{0}$ but $\mathbf{L}^{k-1} \neq \mathbf{0}$. The first question is, "Can $\mathbf{L}$ be diagonalized by a similarity transformation?" To answer this, notice that $\lambda = 0$ is the only eigenvalue of $\mathbf{L}$ because

$$\mathbf{L}\mathbf{x} = \lambda\mathbf{x} \implies \mathbf{L}^k\mathbf{x} = \lambda^k\mathbf{x} \implies \mathbf{0} = \lambda^k\mathbf{x} \implies \lambda = 0 \quad (\text{since } \mathbf{x} \neq \mathbf{0}\,).$$

So if $\mathbf{L}$ is to be diagonalized by a similarity transformation, it must be the case that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{D} = \mathbf{0}$ (diagonal entries of $\mathbf{D}$ must be eigenvalues of $\mathbf{L}$), and this forces $\mathbf{L} = \mathbf{0}$. In other words, the *only* nilpotent matrix that is similar to a diagonal matrix is the zero matrix.

Assume $\mathbf{L} \neq \mathbf{0}$ from now on so that $\mathbf{L}$ is not diagonalizable. Since $\mathbf{L}$ can always be triangularized (Schur's theorem again), our problem boils down to finding a nonsingular $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P}$ is an upper-triangular matrix possessing a simple and predictable form. This turns out to be a fundamental problem, and the rest of this section is devoted to its solution. But before diving in, let's set the stage by thinking about some possibilities.

If $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{T}$ is upper triangular, then the diagonal entries of $\mathbf{T}$ must be the eigenvalues of $\mathbf{L}$, so $\mathbf{T}$ must have the form

$$\mathbf{T} = \begin{pmatrix} 0 & \star & \cdots & \star \\ & \ddots & \ddots & \vdots \\ & & \ddots & \star \\ & & & 0 \end{pmatrix}.$$

One way to simplify the form of $\mathbf{T}$ is to allow nonzero entries only on the superdiagonal (the diagonal immediately above the main diagonal) of $\mathbf{T}$, so we might try to construct a nonsingular $\mathbf{P}$ such that $\mathbf{T}$ has the form

$$\mathbf{T} = \begin{pmatrix} 0 & \star & & \\ & \ddots & \ddots & \\ & & \ddots & \star \\ & & & 0 \end{pmatrix}.$$

To gain some insight on how this might be accomplished, let $\mathbf{L}$ be a 3×3 nilpotent matrix for which $\mathbf{L}^3 = \mathbf{0}$ and $\mathbf{L}^2 \neq \mathbf{0}$, and search for a $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \Longleftrightarrow \mathbf{L}[\,\mathbf{P}_{*1}\,\mathbf{P}_{*2}\,\mathbf{P}_{*3}\,] = [\,\mathbf{P}_{*1}\,\mathbf{P}_{*2}\,\mathbf{P}_{*3}\,]\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\Longleftrightarrow \mathbf{L}\mathbf{P}_{*1} = \mathbf{0}, \quad \mathbf{L}\mathbf{P}_{*2} = \mathbf{P}_{*1}, \quad \mathbf{L}\mathbf{P}_{*3} = \mathbf{P}_{*2}.$$

Since $\mathbf{L}^3 = \mathbf{0}$, we can set $\mathbf{P}_{*1} = \mathbf{L}^2\mathbf{x}$ for any $\mathbf{x}_{3\times 1}$ such that $\mathbf{L}^2\mathbf{x} \neq \mathbf{0}$. This in turn allows us to set $\mathbf{P}_{*2} = \mathbf{L}\mathbf{x}$ and $\mathbf{P}_{*3} = \mathbf{x}$. Because $\mathcal{J} = \{\mathbf{L}^2\mathbf{x}, \mathbf{L}\mathbf{x}, \mathbf{x}\}$ is a linearly independent set (Exercise 5.10.8), $\mathbf{P} = [\,\mathbf{L}^2\mathbf{x}\,|\,\mathbf{L}\mathbf{x}\,|\,\mathbf{x}\,]$ will do the job. $\mathcal{J}$ is called a ***Jordan chain***, and it is characterized by the fact that its first vector is a somewhat special eigenvector for $\mathbf{L}$ while the other vectors are built (or "chained") on top of this eigenvector to form a special basis for $\mathcal{C}^3$. There are a few more wrinkles in the development of a general theory for $n \times n$ nilpotent matrices, but the features illustrated here illuminate the path.

For a general nilpotent matrix $\mathbf{L}_{n\times n} \neq \mathbf{0}$ of index k, we know that $\lambda = 0$ is the only eigenvalue, so the set of eigenvectors of $\mathbf{L}$ is $N(\mathbf{L})$ (excluding the zero vector of course). Realizing that $\mathbf{L}$ is not diagonalizable is equivalent to realizing that $\mathbf{L}$ does not possess a complete linearly independent set of eigenvectors or, equivalently, $\dim N(\mathbf{L}) < n$. As in the 3×3 example above, the strategy for building a similarity transformation $\mathbf{P}$ that reduces $\mathbf{L}$ to a simple triangular form is as follows.

(1) Construct a somewhat special basis $\mathcal{B}$ for $N(\mathbf{L})$.

(2) Extend $\mathcal{B}$ to a basis for $\mathcal{C}^n$ by building Jordan chains on top of the eigenvectors in $\mathcal{B}$.

To accomplish (1), consider the subspaces defined by

$$\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N(\mathbf{L}) \quad \text{for } i = 0, 1, \ldots, k, \tag{7.7.1}$$

and notice (Exercise 7.7.4) that these subspaces are nested as

$$\mathbf{0} = \mathcal{M}_k \subseteq \mathcal{M}_{k-1} \subseteq \mathcal{M}_{k-2} \subseteq \cdots \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_0 = N(\mathbf{L}).$$

Use these nested spaces to construct a basis for $N(\mathbf{L}) = \mathcal{M}_0$ by starting with any basis $\mathcal{S}_{k-1}$ for $\mathcal{M}_{k-1}$ and by sequentially extending $\mathcal{S}_{k-1}$ with additional sets $\mathcal{S}_{k-2}, \mathcal{S}_{k-3}, \ldots, \mathcal{S}_0$ such that $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2}$ is a basis for $\mathcal{M}_{k-2}$, $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \mathcal{S}_{k-3}$ is a basis for $\mathcal{M}_{k-3}$, etc. In general, $\mathcal{S}_i$ is a set of vectors that extends $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_{i-1}$ to a basis for $\mathcal{M}_i$. Figure 7.7.1 is a heuristic diagram depicting an example of $k = 5$ nested subspaces $\mathcal{M}_i$ along with some typical extension sets $\mathcal{S}_i$ that combine to form a basis for $N(\mathbf{L})$.

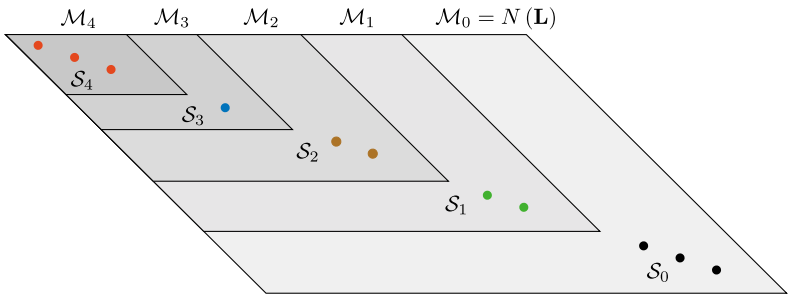


FIGURE 7.7.1

Now extend the basis $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ for $N(\mathbf{L})$ to a basis for $\mathcal{C}^n$ by building Jordan chains on top of each $\mathbf{b} \in \mathcal{B}$. If $\mathbf{b} \in \mathcal{S}_i$, then there exists a vector $\mathbf{x}$ such that $\mathbf{L}^i\mathbf{x} = \mathbf{b}$ because each $\mathbf{b} \in \mathcal{S}_i$ belongs to $\mathcal{M}_i = R(\mathbf{L}^i) \cap N(\mathbf{L}) \subseteq R(\mathbf{L}^i)$. A ***Jordan chain*** is built on top of each $\mathbf{b} \in \mathcal{S}_i$ by solving the system $\mathbf{L}^i\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ and by setting

$$\mathcal{J}_{\mathbf{b}} = \{\mathbf{L}^i\mathbf{x}, \mathbf{L}^{i-1}\mathbf{x}, \ldots, \mathbf{L}\mathbf{x}, \mathbf{x}\}. \tag{7.7.2}$$

Notice that chains built on top of vectors from $\mathcal{S}_i$ each have length $i+1$. The heuristic diagram in Figure 7.7.2 depicts Jordan chains built on top of the basis vectors illustrated in Figure 7.7.1—the chain that is labeled is built on top of a vector $\mathbf{b} \in \mathcal{S}_3$.

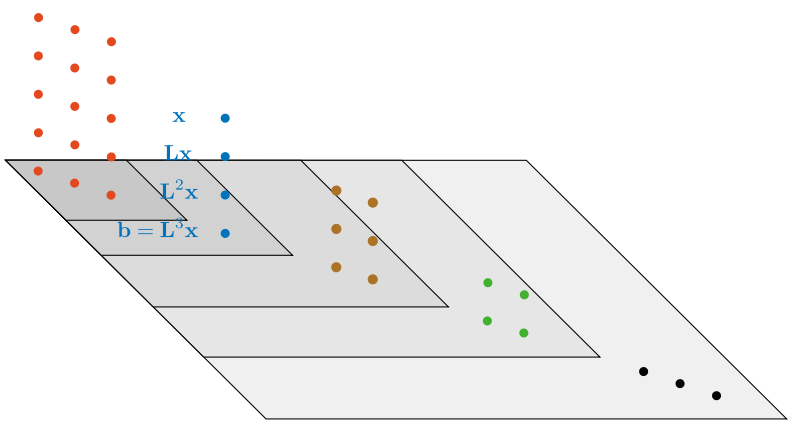


FIGURE 7.7.2

The collection of vectors in all of these Jordan chains is a basis for $\mathcal{C}^n$. To demonstrate this, first it must be argued that the total number of vectors in all Jordan chains is n, and then it must be proven that this collection is a linearly independent set. To count the number of vectors in all Jordan chains $\mathcal{J}_{\mathbf{b}}$, first recall from (4.5.1) that the rank of a product is given by the formula $rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B})$, and apply this to conclude that $\dim \mathcal{M}_i = \dim R\left(\mathbf{L}^i\right) \cap N\,(\mathbf{L}) = rank\left(\mathbf{L}^i\right) - rank\left(\mathbf{L}\mathbf{L}^i\right)$. In other words, if we set $d_i = \dim \mathcal{M}_i$ and $r_i = rank\left(\mathbf{L}^i\right)$, then

$$d_i = \dim \mathcal{M}_i = rank\left(\mathbf{L}^i\right) - rank\left(\mathbf{L}^{i+1}\right) = r_i - r_{i+1}, \tag{7.7.3}$$

so the number of vectors in $\mathcal{S}_i$ is

$$\nu_i = d_i - d_{i+1} = r_i - 2r_{i+1} + r_{i+2}. \tag{7.7.4}$$

Since every chain emanating from a vector in $\mathcal{S}_i$ contains $i+1$ vectors, and since $d_k = 0 = r_k$, the total number of vectors in all Jordan chains is

$$\begin{aligned}
\text{total} &= \sum_{i=0}^{k-1}(i+1)\nu_i = \sum_{i=0}^{k-1}(i+1)(d_i - d_{i+1}) \\
&= d_0 - d_1 + 2(d_1 - d_2) + 3(d_2 - d_3) + \cdots + k(d_{k-1} - d_k) \\
&= d_0 + d_1 + \cdots + d_{k-1} \\
&= (r_0 - r_1) + (r_1 - r_2) + (r_2 - r_3) + \cdots + (r_{k-1} - r_k) \\
&= r_0 = n.
\end{aligned}$$

To prove that the set of all vectors from all Jordan chains is linearly independent, place these vectors as columns in a matrix $\mathbf{Q}_{n\times n}$ and show that $N\,(\mathbf{Q}) = \mathbf{0}$. The trick in doing so is to arrange the vectors from the $\mathcal{J}_{\mathbf{b}}$'s in just the right order. Begin by placing the vectors at the top level in chains emanating from $\mathcal{S}_i$ as columns in a matrix $\mathbf{X}_i$ as depicted in the heuristic diagram in Figure 7.7.3.

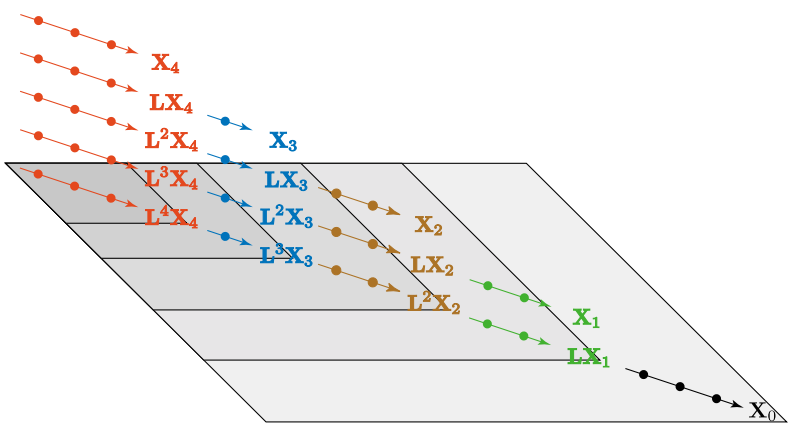


Figure 7.7.3

The matrix $\mathbf{LX}_i$ contains all vectors at the second highest level of those chains emanating from $\mathcal{S}_i$, while $\mathbf{L}^2\mathbf{X}_i$ contains all vectors at the third highest level of those chains emanating from $\mathcal{S}_i$, and so on. In general, $\mathbf{L}^j\mathbf{X}_i$ contains all vectors at the $(j+1)^{st}$ highest level of those chains emanating from $\mathcal{S}_i$. Proceed by filling in $\mathbf{Q} = [\,\mathbf{Q}_0 \,|\, \mathbf{Q}_1 \,|\cdots|\, \mathbf{Q}_{k-1}\,]$ from the bottom up by letting $\mathbf{Q}_j$ be the matrix whose columns are all vectors at the jth level from the bottom in all chains. For the example illustrated in Figures 7.7.1–7.7.3 with $k = 5$,

$$\mathbf{Q}_0 = [\,\mathbf{X}_0 \,|\, \mathbf{LX}_1 \,|\, \mathbf{L}^2\mathbf{X}_2 \,|\, \mathbf{L}^3\mathbf{X}_3 \,|\, \mathbf{L}^4\mathbf{X}_4\,] = \text{vectors at level 0} = \text{basis } \mathcal{B} \text{ for } N(\mathbf{L}),$$

$$\mathbf{Q}_1 = [\,\mathbf{X}_1 \,|\, \mathbf{LX}_2 \,|\, \mathbf{L}^2\mathbf{X}_3 \,|\, \mathbf{L}^3\mathbf{X}^4\,] = \text{ vectors at level 1 (from the bottom)},$$

$$\mathbf{Q}_2 = [\,\mathbf{X}_2 \,|\, \mathbf{LX}_3 \,|\, \mathbf{L}^2\mathbf{X}^4\,] = \text{ vectors at level 2 (from the bottom)},$$

$$\mathbf{Q}_3 = [\,\mathbf{X}_3 \,|\, \mathbf{LX}^4\,] = \text{ vectors at level 3 (from the bottom)},$$

$$\mathbf{Q}_4 = [\,\mathbf{X}_4\,] = \text{ vectors at level 4 (from the bottom)}.$$

In general, $\mathbf{Q}_j = [\,\mathbf{X}_j \,|\, \mathbf{LX}_{j+1} \,|\, \mathbf{L}^2\mathbf{X}_{j+2} \,|\cdots|\, \mathbf{L}^{k-1-j}\mathbf{X}_{k-1}\,]$. Since the columns of $\mathbf{L}^j\mathbf{X}_j$ are all on the bottom level (level 0), they are part of the basis $\mathcal{B}$ for $N(\mathbf{L})$. This means that the columns of $\mathbf{L}^j\mathbf{Q}_j$ are also part of the basis $\mathcal{B}$ for $N(\mathbf{L})$, so they are linearly independent, and thus $N\left(\mathbf{L}^j\mathbf{Q}_j\right) = \mathbf{0}$. Furthermore, since the columns of $\mathbf{L}^j\mathbf{Q}_j$ are in $N(\mathbf{L})$, we have $\mathbf{L}\left(\mathbf{L}^j\mathbf{Q}_j\right) = \mathbf{0}$, and hence $\mathbf{L}^{j+h}\mathbf{Q}_j = \mathbf{0}$ for all $h \geq 1$. Now use these observations to prove $N(\mathbf{Q}) = \mathbf{0}$. If $\mathbf{Qz} = \mathbf{0}$, then multiplication by $\mathbf{L}^{k-1}$ yields

$$\begin{aligned}
\mathbf{0} = \mathbf{L}^{k-1}\mathbf{Qz} &= [\,\mathbf{L}^{k-1}\mathbf{Q}_0 \,|\, \mathbf{L}^{k-1}\mathbf{Q}_1 \,|\cdots|\, \mathbf{L}^{k-1}\mathbf{Q}_{k-1}\,]\,\mathbf{z} \\
&= [\,\mathbf{0} \,|\, \mathbf{0} \,|\cdots|\, \mathbf{L}^{k-1}\mathbf{Q}_{k-1}\,] \begin{pmatrix} \mathbf{z}_0 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_{k-1} \end{pmatrix} \implies \mathbf{z}_{k-1} \in N\left(\mathbf{L}^{k-1}\mathbf{Q}_{k-1}\right) \\
&\implies \mathbf{z}_{k-1} = \mathbf{0}.
\end{aligned}$$

This conclusion with the same argument applied to $\mathbf{0} = \mathbf{L}^{k-2}\mathbf{Qz}$ produces $\mathbf{z}_{k-2} = \mathbf{0}$. Similar repetitions show that $\mathbf{z}_i = \mathbf{0}$ for each i, and thus $N(\mathbf{Q}) = \mathbf{0}$.

It has now been proven that if $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ is the basis for $N(\mathbf{L})$ derived from the nested subspaces $\mathcal{M}_i$, then the set of all Jordan chains $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1} \cup \mathcal{J}_{\mathbf{b}_2} \cup \cdots \cup \mathcal{J}_{\mathbf{b}_t}$ is a basis for $\mathcal{C}^n$. If the vectors from $\mathcal{J}$ are placed as columns (in the order in which they appear in $\mathcal{J}$) in a matrix $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1 \,|\, \mathbf{J}_2 \,|\cdots|\, \mathbf{J}_t\,]$, then $\mathbf{P}$ is nonsingular, and if $\mathbf{b}_j \in \mathcal{S}_i$, then $\mathbf{J}_j = [\,\mathbf{L}^i\mathbf{x} \,|\, \mathbf{L}^{i-1}\mathbf{x} \,|\cdots|\, \mathbf{Lx} \,|\, \mathbf{x}\,]$ for some $\mathbf{x}$ such that $\mathbf{L}^i\mathbf{x} = \mathbf{b}_j$ so that

$$\mathbf{LJ}_j = [\,\mathbf{0} \,|\, \mathbf{L}^i\mathbf{x} \,|\cdots|\, \mathbf{Lx}\,] = [\,\mathbf{L}^i\mathbf{x} \,|\cdots|\, \mathbf{Lx} \,|\, \mathbf{x}\,] \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} = \mathbf{J}_j\mathbf{N}_j,$$

where $\mathbf{N}_j$ is an $(i+1)\times(i+1)$ matrix whose entries are equal to 1 along the superdiagonal and zero elsewhere. Therefore,

$$\mathbf{LP} = [\,\mathbf{LJ}_1\,|\,\mathbf{LJ}_2\,|\cdots|\,\mathbf{LJ}_t\,] = [\,\mathbf{J}_1\,|\,\mathbf{J}_2\,|\cdots|\,\mathbf{J}_t\,]\begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix}$$

or, equivalently,

$$\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix}, \text{ where } \mathbf{N}_j = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}. \tag{7.7.5}$$

Each $\mathbf{N}_j$ is a nilpotent matrix whose index is given by its size. The $\mathbf{N}_j$'s are called ***nilpotent Jordan blocks***, and the block-diagonal matrix $\mathbf{N}$ is called the ***Jordan form*** for $\mathbf{L}$. Below is a summary.

Jordan Form for a Nilpotent Matrix

Every nilpotent matrix $\mathbf{L}_{n\times n}$ of index k is similar to a block-diagonal matrix

$$\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix} \tag{7.7.6}$$

in which each $\mathbf{N}_j$ is a nilpotent matrix having ones on the superdiagonal and zeros elsewhere—see (7.7.5).

- The number of blocks in $\mathbf{N}$ is given by $t = \dim N(\mathbf{L})$.
- The size of the largest block in $\mathbf{N}$ is $k\times k$.
- The number of $i\times i$ blocks in $\mathbf{N}$ is $\nu_i = r_{i-1} - 2r_i + r_{i+1}$, where $r_i = rank\left(\mathbf{L}^i\right)$—this follows from (7.7.4).
- If $\mathcal{B} = \mathcal{S}_{k-1}\cup\mathcal{S}_{k-2}\cup\cdots\cup\mathcal{S}_0 = \{\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_t\}$ is a basis for $N(\mathbf{L})$ derived from the nested subspaces $\mathcal{M}_i = R\left(\mathbf{L}^i\right)\cap N(\mathbf{L})$, then
  - ▷ the set of vectors $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1}\cup\mathcal{J}_{\mathbf{b}_2}\cup\cdots\cup\mathcal{J}_{\mathbf{b}_t}$ from all Jordan chains is a basis for $\mathcal{C}^n$;
  - ▷ $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1\,|\,\mathbf{J}_2\,|\cdots|\,\mathbf{J}_t\,]$ is the nonsingular matrix containing these Jordan chains in the order in which they appear in $\mathcal{J}$.

The following theorem demonstrates that the ***Jordan structure*** (the number and the size of the blocks in $\mathbf{N}$) is uniquely determined by $\mathbf{L}$, but $\mathbf{P}$ is not. In other words, the Jordan form is unique up to the arrangement of the individual Jordan blocks.

Uniqueness of the Jordan Structure

The structure of the Jordan form for a nilpotent matrix $\mathbf{L}_{n\times n}$ of index k is uniquely determined by $\mathbf{L}$ in the sense that whenever $\mathbf{L}$ is similar to a block-diagonal matrix $\mathbf{B} = \operatorname{diag}(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_t)$ in which each $\mathbf{B}_i$ has the form

$$\mathbf{B}_i = \begin{pmatrix} 0 & \epsilon_i & 0 & \cdots & 0 \\ 0 & 0 & \epsilon_i & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \epsilon_i \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}_{n_i\times n_i} \qquad \text{for} \quad \epsilon_i \neq 0,$$

then it must be the case that $t = \dim N(\mathbf{L})$, and the number of blocks having size $i \times i$ must be given by $r_{i-1} - 2r_i + r_{i+1}$, where $r_i = rank\left(\mathbf{L}^i\right)$.

Proof. Suppose that $\mathbf{L}$ is similar to both $\mathbf{B}$ and $\mathbf{N}$, where $\mathbf{B}$ is as described above and $\mathbf{N}$ is as described in (7.7.6). This implies that $\mathbf{B}$ and $\mathbf{N}$ are similar, and hence $rank\left(\mathbf{B}^i\right) = rank\left(\mathbf{L}^i\right) = r_i$ for every nonnegative integer i. In particular, $index(\mathbf{B}) = index(\mathbf{L})$. Each time a block $\mathbf{B}_i$ is powered, the line of ϵ_i's moves to the next higher diagonal level so that

$$rank\left(\mathbf{B}_i^p\right) = \begin{cases} n_i - p & \text{if } p < n_i, \\ 0 & \text{if } p \geq n_i. \end{cases}$$

Since $r_p = rank\left(\mathbf{B}^p\right) = \sum_{i=1}^{t} rank\left(\mathbf{B}_i^p\right)$, it follows that if ω_i is the number of $i \times i$ blocks in $\mathbf{B}$, then

$$\begin{aligned} r_{k-1} &= \omega_k, \\ r_{k-2} &= \omega_{k-1} + 2\omega_k, \\ r_{k-3} &= \omega_{k-2} + 2\omega_{k-1} + 3\omega_k, \\ &\vdots \end{aligned}$$

and, in general, $r_i = \omega_{i+1} + 2\omega_{i+2} + \cdots + (k-i)\omega_k$. It's now straightforward to verify that $r_{i-1} - 2r_i + r_{i+1} = \omega_i$. Finally, using this equation together with (7.7.4) guarantees that the number of blocks in $\mathbf{B}$ must be

$$t = \sum_{i=1}^{k} \omega_i = \sum_{i=1}^{k} (r_{i-1} - 2r_i + r_{i+1}) = \sum_{i=1}^{k} \nu_i = \dim N(\mathbf{L}). \quad \blacksquare$$

The manner in which we developed the Jordan theory spawned 1's on the superdiagonals of the Jordan blocks $\mathbf{N}_i$ in (7.7.5). But it was not necessary to do so—it was simply a matter of convenience. In fact, any nonzero value can be forced onto the superdiagonal of any $\mathbf{N}_i$—see Exercise 7.7.9. In other words, the fact that 1's appear on the superdiagonals of the $\mathbf{N}_i$'s is artificial and is not important to the structure of the Jordan form for $\mathbf{L}$. What's important, and what constitutes the "Jordan structure," is the number and sizes of the Jordan blocks (or chains) and not the values appearing on the superdiagonals of these blocks.

Example 7.7.1

Problem: Determine the Jordan forms for 3×3 nilpotent matrices $\mathbf{L}_1$, $\mathbf{L}_2$, and $\mathbf{L}_3$ that have respective indices $k = 1, 2, 3$.

Solution: The size of the largest block must be $k \times k$, so

$$\mathbf{N}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{N}_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{N}_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Example 7.7.2

For a nilpotent matrix $\mathbf{L}$, the theoretical development relies on a complicated basis for $N(\mathbf{L})$ to derive the structure of the Jordan form $\mathbf{N}$ as well as the Jordan chains that constitute a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$. But, after the dust settled, we saw that a basis for $N(\mathbf{L})$ is not needed to construct $\mathbf{N}$ because $\mathbf{N}$ is completely determined simply by ranks of powers of $\mathbf{L}$. A basis for $N(\mathbf{L})$ is only required to construct the Jordan chains in $\mathbf{P}$.

Question: For the purpose of constructing Jordan chains in $\mathbf{P}$, can we use an arbitrary basis for $N(\mathbf{L})$ instead of the complicated basis built from the $\mathcal{M}_i$'s?

Answer: No! Consider the nilpotent matrix

$$\mathbf{L} = \begin{pmatrix} 2 & 0 & 1 \\ -4 & 0 & -2 \\ -4 & 0 & -2 \end{pmatrix} \quad \text{and its Jordan form} \quad \mathbf{N} = \left(\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right).$$

If $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$, where $\mathbf{P} = [\,\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3\,]$, then $\mathbf{L}\mathbf{P} = \mathbf{P}\mathbf{N}$ implies that $\mathbf{L}\mathbf{x}_1 = \mathbf{0}$, $\mathbf{L}\mathbf{x}_2 = \mathbf{x}_1$, and $\mathbf{L}\mathbf{x}_3 = \mathbf{0}$. In other words, $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_3\}$ must be a basis for $N(\mathbf{L})$, and $\mathcal{J}_{\mathbf{x}_1} = \{\mathbf{x}_1, \mathbf{x}_2\}$ must be a Jordan chain built on top of $\mathbf{x}_1$. If we try to construct such vectors by starting with the naive basis

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{7.7.7}$$

for $N(\mathbf{L})$ obtained by solving $\mathbf{Lx} = \mathbf{0}$ with straightforward Gaussian elimination, we immediately hit a brick wall because $\mathbf{x}_1 \notin R(\mathbf{L})$ means $\mathbf{Lx}_2 = \mathbf{x}_1$ is an inconsistent system, so $\mathbf{x}_2$ cannot be determined. Similarly, $\mathbf{x}_3 \notin R(\mathbf{L})$ insures that the same difficulty occurs if $\mathbf{x}_3$ is used in place of $\mathbf{x}_1$. In other words, even though the vectors in (7.7.7) constitute an otherwise perfectly good basis for $N(\mathbf{L})$, they can't be used to build $\mathbf{P}$.

Example 7.7.3

Problem: Let $\mathbf{L}_{n\times n}$ be a nilpotent matrix of index k. Provide an algorithm for constructing the Jordan chains that generate a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form.

Solution:

1. Start with the fact that $\mathcal{M}_{k-1} = R\left(\mathbf{L}^{k-1}\right)$ (Exercise 7.7.5), and determine a basis $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$ for $R\left(\mathbf{L}^{k-1}\right)$.
2. Extend $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$ to a basis for $\mathcal{M}_{k-2} = R\left(\mathbf{L}^{k-2}\right) \cap N(\mathbf{L})$ as follows.
   - ▷ Find a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ for $N(\mathbf{LB})$, where $\mathbf{B}$ is a matrix containing a basis for $R\left(\mathbf{L}^{k-2}\right)$—e.g., the basic columns of $\mathbf{L}^{k-2}$. The set $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \ldots, \mathbf{Bv}_s\}$ is a basis for $\mathcal{M}_{k-2}$ (see p. 211).
   - ▷ Find the basic columns in $[\,\mathbf{y}_1 \,|\, \mathbf{y}_2 \,|\, \cdots \,|\, \mathbf{y}_q \,|\, \mathbf{Bv}_1 \,|\, \mathbf{Bv}_2 \,|\, \cdots \,|\, \mathbf{Bv}_s\,]$. Say they are $\{\mathbf{y}_1, \ldots, \mathbf{y}_q, \mathbf{Bv}_{\beta_1}, \ldots, \mathbf{Bv}_{\beta_j}\}$ (all of the $\mathbf{y}_j$'s are basic because they are a leading linearly independent subset). This is a basis for $\mathcal{M}_{k-2}$ that contains a basis for $\mathcal{M}_{k-1}$. In other words,

$$\mathcal{S}_{k-1} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\} \quad \text{and} \quad \mathcal{S}_{k-2} = \{\mathbf{Bv}_{\beta_1}, \mathbf{Bv}_{\beta_2}, \ldots, \mathbf{Bv}_{\beta_j}\}.$$

3. Repeat the above procedure $k-1$ times to construct a basis for $N(\mathbf{L})$ that is of the form $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$, where $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_i$ is a basis for $\mathcal{M}_i$ for each $i = k-1, k-2, \ldots, 0$.
4. Build a Jordan chain on top of each $\mathbf{b}_j \in \mathcal{B}$. If $\mathbf{b}_j \in \mathcal{S}_i$, then we solve $\mathbf{L}^i\mathbf{x}_j = \mathbf{b}_j$ and set $\mathbf{J}_j = [\,\mathbf{L}^i\mathbf{x}_j \,|\, \mathbf{L}^{i-1}\mathbf{x}_j \,|\, \cdots \,|\, \mathbf{Lx}_j \,|\, \mathbf{x}_j\,]$. The desired similarity transformation is $\mathbf{P}_{n\times n} = [\,\mathbf{J}_1 \,|\, \mathbf{J}_2 \,|\, \cdots \,|\, \mathbf{J}_t\,]$.

Example 7.7.4

Problem: Find $\mathbf{P}$ and $\mathbf{N}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form, where

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

Solution: First determine the Jordan form for $\mathbf{L}$. Computing $r_i = rank\left(\mathbf{L}^i\right)$ reveals that $r_1 = 3$, $r_2 = 1$, and $r_3 = 0$, so the index of $\mathbf{L}$ is $k = 3$, and

$$\begin{aligned}
\text{the number of } 3\times 3 \text{ blocks } &= r_2 - 2r_3 + r_4 = 1,\\
\text{the number of } 2\times 2 \text{ blocks } &= r_1 - 2r_2 + r_3 = 1,\\
\text{the number of } 1\times 1 \text{ blocks } &= r_0 - 2r_1 + r_2 = 1.
\end{aligned}$$

Consequently, the Jordan form of $\mathbf{L}$ is

$$\mathbf{N} = \left(\begin{array}{ccc|cc|c}
0 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0\\
\hline
0 & 0 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 0 & 0 & 0\\
\hline
0 & 0 & 0 & 0 & 0 & 0
\end{array}\right).$$

Notice that three Jordan blocks were found, and this agrees with the fact that $\dim N(\mathbf{L}) = 6 - rank(\mathbf{L}) = 3$. Determine $\mathbf{P}$ by following the procedure described in Example 7.7.3.

1. Since $rank\left(\mathbf{L}^2\right) = 1$, any nonzero column from $\mathbf{L}^2$ will be a basis for $\mathcal{M}_2 = R\left(\mathbf{L}^2\right)$, so set $\mathbf{y}_1 = [\mathbf{L}^2]_{*1} = (6,-6,0,0,-6,-6)^T$.

2. To extend $\mathbf{y}_1$ to a basis for $\mathcal{M}_1 = R(\mathbf{L}) \cap N(\mathbf{L})$, use

$$\mathbf{B} = [\mathbf{L}_{*1}\,|\,\mathbf{L}_{*2}\,|\,\mathbf{L}_{*3}] = \begin{pmatrix} 1 & 1 & -2\\ 3 & 1 & 5\\ -2 & -1 & 0\\ 2 & 1 & 0\\ -5 & -3 & -1\\ -3 & -2 & -1 \end{pmatrix} \implies \mathbf{LB} = \begin{pmatrix} 6 & 3 & 3\\ -6 & -3 & -3\\ 0 & 0 & 0\\ 0 & 0 & 0\\ -6 & -3 & -3\\ -6 & -3 & -3 \end{pmatrix},$$

and determine a basis for $N(\mathbf{LB})$ to be $\left\{\mathbf{v}_1 = \begin{pmatrix} -1\\ 2\\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -1\\ 0\\ 2 \end{pmatrix}\right\}$. Reducing $[\,\mathbf{y}_1\,|\,\mathbf{Bv}_1\,|\,\mathbf{Bv}_2\,]$ to echelon form shows that its basic columns are in the first and third positions, so $\{\mathbf{y}_1, \mathbf{Bv}_2\}$ is a basis for $\mathcal{M}_1$ with

$$\mathcal{S}_2 = \left\{\begin{pmatrix} 6\\ -6\\ 0\\ 0\\ -6\\ -6 \end{pmatrix} = \mathbf{b}_1\right\} \quad \text{and} \quad \mathcal{S}_1 = \left\{\begin{pmatrix} -5\\ 7\\ 2\\ -2\\ 3\\ 1 \end{pmatrix} = \mathbf{b}_2\right\}.$$

3. Now extend $\mathcal{S}_2 \cup \mathcal{S}_1 = \{\mathbf{b}_1, \mathbf{b}_2\}$ to a basis for $\mathcal{M}_0 = N(\mathbf{L})$. This time, $\mathbf{B} = \mathbf{I}$, and a basis for $N(\mathbf{LB}) = N(\mathbf{L})$ can be computed to be

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -4 \\ -1 \\ 3 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}_2 = \begin{pmatrix} -4 \\ 5 \\ 2 \\ 0 \\ 3 \\ 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ -2 \\ 0 \\ 0 \\ 3 \end{pmatrix},$$

and $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \mathbf{Bv}_3\} = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. Reducing $[\,\mathbf{b}_1 \,|\, \mathbf{b}_2 \,|\, \mathbf{v}_1 \,|\, \mathbf{v}_2 \,|\, \mathbf{v}_3\,]$ to echelon form reveals that its basic columns are in positions one, two, and three, so $\mathbf{v}_1$ is the needed extension vector. Therefore, the complete nested basis for $N(\mathbf{L})$ is

$$\mathbf{b}_1 = \begin{pmatrix} 6 \\ -6 \\ 0 \\ 0 \\ -6 \\ -6 \end{pmatrix} \in \mathcal{S}_2, \quad \mathbf{b}_2 = \begin{pmatrix} -5 \\ 7 \\ 2 \\ -2 \\ 3 \\ 1 \end{pmatrix} \in \mathcal{S}_1, \quad \text{and} \quad \mathbf{b}_3 = \begin{pmatrix} 2 \\ -4 \\ -1 \\ 3 \\ 0 \\ 0 \end{pmatrix} \in \mathcal{S}_0.$$

4. Complete the process by building a Jordan chain on top of each $\mathbf{b}_j \in \mathcal{S}_i$ by solving $\mathbf{L}^i\mathbf{x}_j = \mathbf{b}_j$ and by setting $\mathbf{J}_j = [\mathbf{L}^i\mathbf{x}_j \,|\, \cdots \,|\, \mathbf{Lx}_j \,|\, \mathbf{x}_j\,]$. Since $\mathbf{x}_1 = \mathbf{e}_1$ solves $\mathbf{L}^2\mathbf{x}_1 = \mathbf{b}_1$, we have $\mathbf{J}_1 = [\,\mathbf{L}^2\mathbf{e}_1 \,|\, \mathbf{Le}_1 \,|\, \mathbf{e}_1\,]$. Solving $\mathbf{Lx}_2 = \mathbf{b}_2$ yields $\mathbf{x}_2 = (-1, 0, 2, 0, 0, 0)^T$, so $\mathbf{J}_2 = [\,\mathbf{Lx}_2 \,|\, \mathbf{x}_2\,]$. Finally, $\mathbf{J}_3 = [\,\mathbf{b}_3\,]$. Putting these chains together produces

$$\mathbf{P} = [\,\mathbf{J}_1 \,|\, \mathbf{J}_2 \,|\, \mathbf{J}_3\,] = \begin{pmatrix} 6 & 1 & 1 & -5 & -1 & 2 \\ -6 & 3 & 0 & 7 & 0 & -4 \\ 0 & -2 & 0 & 2 & 2 & -1 \\ 0 & 2 & 0 & -2 & 0 & 3 \\ -6 & -5 & 0 & 3 & 0 & 0 \\ -6 & -3 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

It can be verified by direct multiplication that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$.

It's worthwhile to pay attention to how the results in this section translate into the language of direct sum decompositions of invariant subspaces as discussed in §4.9 (p. 259) and §5.9 (p. 383). For a linear nilpotent operator $\mathbf{L}$ of index k defined on a finite-dimensional vector space $\mathcal{V}$, statement (7.7.6) on p. 579 means that $\mathcal{V}$ can be decomposed as a direct sum $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \cdots \oplus \mathcal{V}_t$, where $\mathcal{V}_j = \operatorname{span}(\mathcal{J}_{\mathbf{b}_j})$ is the space spanned by a Jordan chain emanating from the basis vector $\mathbf{b}_j \in N(\mathbf{L})$ and where $t = \dim N(\mathbf{L})$. Furthermore, each $\mathcal{V}_j$ is an

invariant subspace for $\mathbf{L}$, and the matrix representation of $\mathbf{L}$ with respect to the basis $\mathcal{J} = \mathcal{J}_{\mathbf{b}_1} \cup \mathcal{J}_{\mathbf{b}_2} \cup \cdots \cup \mathcal{J}_{\mathbf{b}_t}$ is

$$[\mathbf{L}]_{\mathcal{J}} = \begin{pmatrix} \mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_t \end{pmatrix} \quad \text{in which} \quad \mathbf{N}_j = \left[\mathbf{L}_{/\mathcal{V}_j}\right]_{\mathcal{J}_{\mathbf{b}_j}}. \tag{7.7.8}$$

Exercises for section 7.7

7.7.1. Can the index of an $n \times n$ nilpotent matrix ever exceed n?

7.7.2. Determine all possible Jordan forms $\mathbf{N}$ for a 4×4 nilpotent matrix.

7.7.3. Explain why the number of blocks of size $i \times i$ or larger in the Jordan form for a nilpotent matrix is given by $rank\left(\mathbf{L}^{i-1}\right) - rank\left(\mathbf{L}^{i}\right)$.

7.7.4. For a nilpotent matrix $\mathbf{L}$ of index k, let $\mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N\left(\mathbf{L}\right)$. Prove that $\mathcal{M}_i \subseteq \mathcal{M}_{i-1}$ for each $i = 0, 1, \ldots, k$.

7.7.5. Prove that $R\left(\mathbf{L}^{k-1}\right) \cap N\left(\mathbf{L}\right) = R\left(\mathbf{L}^{k-1}\right)$ for all nilpotent matrices $\mathbf{L}$ of index $k > 1$. In other words, prove $\mathcal{M}_{k-1} = R\left(\mathbf{L}^{k-1}\right)$.

7.7.6. Let $\mathbf{L}$ be a nilpotent matrix of index $k > 1$. Prove that if the columns of $\mathbf{B}$ are a basis for $R\left(\mathbf{L}^i\right)$ for $i \leq k-1$, and if $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s\}$ is a basis for $N\left(\mathbf{LB}\right)$, then $\{\mathbf{Bv}_1, \mathbf{Bv}_2, \ldots, \mathbf{Bv}_s\}$ is a basis for $\mathcal{M}_i$.

7.7.7. Find $\mathbf{P}$ and $\mathbf{N}$ such that $\mathbf{P}^{-1}\mathbf{LP} = \mathbf{N}$ is in Jordan form, where

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

7.7.8. Determine the Jordan form for the following 8×8 nilpotent matrix.

$$\mathbf{L} = \begin{pmatrix} 41 & 30 & 15 & 7 & 4 & 6 & 1 & 3 \\ -54 & -39 & -19 & -9 & -6 & -8 & -2 & -4 \\ 9 & 6 & 2 & 1 & 2 & 1 & 0 & 1 \\ -6 & -5 & -3 & -2 & 1 & -1 & 0 & 0 \\ -32 & -24 & -13 & -6 & -2 & -5 & -1 & -2 \\ -10 & -7 & -2 & 0 & -3 & 0 & 3 & -2 \\ -4 & -3 & -2 & -1 & 0 & -1 & -1 & 0 \\ 17 & 12 & 6 & 3 & 2 & 3 & 2 & 1 \end{pmatrix}.$$

7.7.9. Prove that if $\mathbf{N}$ is the Jordan form for a nilpotent matrix $\mathbf{L}$ as described in (7.7.5) and (7.7.6) on p. 579, then for any set of nonzero scalars $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_t\}$, the matrix $\mathbf{L}$ is similar to a matrix $\tilde{\mathbf{N}}$ of the form

$$\tilde{\mathbf{N}} = \begin{pmatrix} \epsilon_1\mathbf{N}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \epsilon_2\mathbf{N}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \epsilon_t\mathbf{N}_t \end{pmatrix}.$$

In other words, the 1's on the superdiagonal of the $\mathbf{N}_i$'s in (7.7.5) are artificial because any nonzero value can be forced onto the superdiagonal of any $\mathbf{N}_i$. What's important in the "Jordan structure" of $\mathbf{L}$ is the number and sizes of the nilpotent Jordan blocks (or chains) and not the values appearing on the superdiagonals of these blocks.

7.8 JORDAN FORM

The goal of this section is to do for general matrices $\mathbf{A} \in \mathcal{C}^{n\times n}$ what was done for nilpotent matrices in §7.7—reduce $\mathbf{A}$ by means of a similarity transformation to a block-diagonal matrix in which each block has a simple triangular form. The two major components for doing this are now in place—they are the core-nilpotent decomposition (p. 397) and the Jordan form for nilpotent matrices. All that remains is to connect these two ideas. To do so, it is convenient to adopt the following terminology.

Index of an Eigenvalue

The ***index of an eigenvalue*** λ for a matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the index of the matrix $(\mathbf{A} - \lambda\mathbf{I})$. In other words, from the characterizations of index given on p. 395, $index(\lambda)$ is the smallest positive integer k such that any one of the following statements is true.

- $rank\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = rank\big((\mathbf{A} - \lambda\mathbf{I})^{k+1}\big)$.
- $R\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = R\big((\mathbf{A} - \lambda\mathbf{I})^{k+1}\big)$.
- $N\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = N\big((\mathbf{A} - \lambda\mathbf{I})^{k+1}\big)$.
- $R\big((\mathbf{A} - \lambda\mathbf{I})^k\big) \cap N\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = \mathbf{0}$.
- $\mathcal{C}^n = R\big((\mathbf{A} - \lambda\mathbf{I})^k\big) \oplus N\big((\mathbf{A} - \lambda\mathbf{I})^k\big)$.

It is understood that $index(\mu) = 0$ if and only if $\mu \notin \sigma(\mathbf{A})$.

The Jordan form for $\mathbf{A} \in \mathcal{C}^{n\times n}$ is derived by digesting the distinct eigenvalues in $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ one at a time with a core-nilpotent decomposition as follows. If $index(\lambda_1) = k_1$, then there is a nonsingular matrix $\mathbf{X}_1$ such that

$$\mathbf{X}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{X}_1 = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{pmatrix}, \tag{7.8.1}$$

where $\mathbf{L}_1$ is nilpotent of index k_1 and $\mathbf{C}_1$ is nonsingular (it doesn't matter whether $\mathbf{C}_1$ or $\mathbf{L}_1$ is listed first, so, for the sake of convenience, the nilpotent block is listed first). We know from the results on nilpotent matrices (p. 579) that there is a nonsingular matrix $\mathbf{Y}_1$ such that

$$\mathbf{Y}_1^{-1}\mathbf{L}_1\mathbf{Y}_1 = \mathbf{N}(\lambda_1) = \begin{pmatrix} \mathbf{N}_1(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{N}_2(\lambda_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{N}_{t_1}(\lambda_1) \end{pmatrix}$$

is a block-diagonal matrix that is characterized by the following features.

▷ Every block in $\mathbf{N}(\lambda_1)$ has the form $\mathbf{N}_\star(\lambda_1) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}$.

▷ There are $t_1 = \dim N(\mathbf{L}_1) = \dim N(\mathbf{A} - \lambda_1\mathbf{I})$ such blocks in $\mathbf{N}(\lambda_1)$.

▷ The number of $i \times i$ blocks of the form $\mathbf{N}_\star(\lambda_1)$ contained in $\mathbf{N}(\lambda_1)$ is $\nu_i(\lambda_1) = rank\left(\mathbf{L}_1^{i-1}\right) - 2\,rank\left(\mathbf{L}_1^{i}\right) + rank\left(\mathbf{L}_1^{i+1}\right)$. But $\mathbf{C}_1$ in (7.8.1) is nonsingular, so $rank\left(\mathbf{L}_1^{p}\right) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^p\right) - rank\left(\mathbf{C}_1\right)$, and thus the number of $i \times i$ blocks $\mathbf{N}_\star(\lambda_1)$ contained in $\mathbf{N}(\lambda_1)$ can be expressed as

$$\nu_i(\lambda_1) = r_{i-1}(\lambda_1) - 2r_i(\lambda_1) + r_{i+1}(\lambda_1), \quad \text{where} \quad r_i(\lambda_1) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^i\right).$$

Now, $\mathbf{Q}_1 = \mathbf{X}_1\begin{pmatrix} \mathbf{Y}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ is nonsingular, and $\mathbf{Q}_1^{-1}(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{Q}_1 = \begin{pmatrix} \mathbf{N}(\lambda_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 \end{pmatrix}$ or, equivalently,

$$\mathbf{Q}_1^{-1}\mathbf{A}\mathbf{Q}_1 = \begin{pmatrix} \mathbf{N}(\lambda_1) + \lambda_1\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_1 + \lambda_1\mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_1 \end{pmatrix}. \tag{7.8.2}$$

The upper-left-hand segment $\mathbf{J}(\lambda_1) = \mathbf{N}(\lambda_1) + \lambda_1\mathbf{I}$ has the block-diagonal form

$$\mathbf{J}(\lambda_1) = \begin{pmatrix} \mathbf{J}_1(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_1) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_1}(\lambda_1) \end{pmatrix} \quad \text{with} \quad \mathbf{J}_\star(\lambda_1) = \mathbf{N}_\star(\lambda_1) + \lambda_1\mathbf{I}.$$

The matrix $\mathbf{J}(\lambda_1)$ is called the ***Jordan segment*** associated with the eigenvalue λ_1, and the individual blocks $\mathbf{J}_\star(\lambda_1)$ contained in $\mathbf{J}(\lambda_1)$ are called ***Jordan blocks*** associated with the eigenvalue λ_1. The structure of the Jordan segment $\mathbf{J}(\lambda_1)$ is inherited from Jordan structure of the associated nilpotent matrix $\mathbf{L}_1$.

▷ Each Jordan block looks like $\mathbf{J}_\star(\lambda_1) = \mathbf{N}_\star(\lambda_1) + \lambda_1\mathbf{I} = \begin{pmatrix} \lambda_1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_1 \end{pmatrix}$.

▷ There are $t_1 = \dim N(\mathbf{A} - \lambda_1\mathbf{I})$ such Jordan blocks in the segment $\mathbf{J}(\lambda_1)$.

▷ The number of $i \times i$ Jordan blocks $\mathbf{J}_\star(\lambda_1)$ contained in $\mathbf{J}(\lambda_1)$ is

$$\nu_i(\lambda_1) = r_{i-1}(\lambda_1) - 2r_i(\lambda_1) + r_{i+1}(\lambda_1), \quad \text{where} \quad r_i(\lambda_1) = rank\left((\mathbf{A} - \lambda_1\mathbf{I})^i\right).$$

Since the distinct eigenvalues of $\mathbf{A}$ are $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, the distinct eigenvalues of $\mathbf{A} - \lambda_1\mathbf{I}$ are

$$\sigma(\mathbf{A} - \lambda_1\mathbf{I}) = \{0,\ (\lambda_2 - \lambda_1),\ (\lambda_3 - \lambda_1), \ldots, (\lambda_s - \lambda_1)\}.$$

Couple this with the fact that the only eigenvalue for the nilpotent matrix $\mathbf{L}_1$ in (7.8.1) is zero to conclude that

$$\sigma(\mathbf{C}_1) = \{(\lambda_2 - \lambda_1),\ (\lambda_3 - \lambda_1), \ldots, (\lambda_s - \lambda_1)\}.$$

Therefore, the spectrum of $\mathbf{A}_1 = \mathbf{C}_1 + \lambda_1\mathbf{I}$ in (7.8.2) is $\sigma(\mathbf{A}_1) = \{\lambda_2, \lambda_3, \ldots, \lambda_s\}$. This means that the core-nilpotent decomposition process described above can be repeated on $\mathbf{A}_1 - \lambda_2\mathbf{I}$ to produce a nonsingular matrix $\mathbf{Q}_2$ such that

$$\mathbf{Q}_2^{-1}\mathbf{A}_1\mathbf{Q}_2 = \begin{pmatrix} \mathbf{J}(\lambda_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \sigma(\mathbf{A}_2) = \{\lambda_3,\ \lambda_4, \ldots, \lambda_s\}, \tag{7.8.3}$$

and where $\mathbf{J}(\lambda_2) = \operatorname{diag}(\mathbf{J}_1(\lambda_2), \mathbf{J}_2(\lambda_2), \ldots, \mathbf{J}_{t_2}(\lambda_2))$ is a Jordan segment composed of Jordan blocks $\mathbf{J}_\star(\lambda_2)$ with the following characteristics.

- ▷ Each Jordan block in $\mathbf{J}(\lambda_2)$ has the form $\mathbf{J}_\star(\lambda_2) = \begin{pmatrix} \lambda_2 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_2 \end{pmatrix}$.
- ▷ There are $t_2 = \dim N(\mathbf{A} - \lambda_2\mathbf{I})$ Jordan blocks in segment $\mathbf{J}(\lambda_2)$.
- ▷ The number of $i \times i$ Jordan blocks in segment $\mathbf{J}(\lambda_2)$ is $\nu_i(\lambda_2) = r_{i-1}(\lambda_2) - 2r_i(\lambda_2) + r_{i+1}(\lambda_2)$, where $r_i(\lambda_2) = rank\big((\mathbf{A} - \lambda_2\mathbf{I})^i\big)$.

If we set $\mathbf{P}_2 = \mathbf{Q}_1\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$, then $\mathbf{P}_2$ is a nonsingular matrix such that

$$\mathbf{P}_2^{-1}\mathbf{A}\mathbf{P}_2 = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(\lambda_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \sigma(\mathbf{A}_2) = \{\lambda_3,\ \lambda_4, \ldots, \lambda_s\}.$$

Repeating this process until all eigenvalues have been depleted results in a nonsingular matrix $\mathbf{P}_s$ such that $\mathbf{P}_s^{-1}\mathbf{A}\mathbf{P}_s = \mathbf{J} = \operatorname{diag}(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s))$ in which each $\mathbf{J}(\lambda_j)$ is a Jordan segment containing $t_j = \dim N(\mathbf{A} - \lambda_j\mathbf{I})$ Jordan blocks. The matrix $\mathbf{J}$ is called the ***Jordan form***[79] for $\mathbf{A}$ (some texts refer to $\mathbf{J}$ as the Jordan *canonical* form or the Jordan *normal* form). The ***Jordan structure*** of $\mathbf{A}$ is defined to be the number of Jordan segments in $\mathbf{J}$ along with the number and sizes of the Jordan blocks within each segment. The proof of uniqueness of the Jordan form for a nilpotent matrix (p. 580) can be extended to all $\mathbf{A} \in \mathcal{C}^{n\times n}$. In other words, the Jordan structure of a matrix is uniquely determined by its entries. Below is a formal summary of these developments.

---

[79] Marie Ennemond Camille Jordan (1838–1922) discussed this idea (not over the complex numbers but over a finite field) in 1870 in *Traité des substitutions et des équations algebraique* that earned him the Poncelet Prize of the Académie des Science. But Jordan may not have been the first to develop these concepts. It has been reported that the German mathematician Karl Theodor Wilhelm Weierstrass (1815–1897) had previously formulated results along these lines. However, Weierstrass did not publish his ideas because he was fanatical about rigor, and he would not release his work until he was sure it was on a firm mathematical foundation. Weierstrass once said that "a mathematician who is not also something of a poet will never be a perfect mathematician."

Jordan Form

For every $\mathbf{A} \in \mathcal{C}^{n\times n}$ with distinct eigenvalues $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, there is a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \begin{pmatrix} \mathbf{J}(\lambda_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(\lambda_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}(\lambda_s) \end{pmatrix}. \tag{7.8.4}$$

- $\mathbf{J}$ has one *Jordan segment* $\mathbf{J}(\lambda_j)$ for each eigenvalue $\lambda_j \in \sigma(\mathbf{A})$.
- Each segment $\mathbf{J}(\lambda_j)$ is made up of $t_j = \dim N(\mathbf{A} - \lambda_j\mathbf{I})$ *Jordan blocks* $\mathbf{J}_\star(\lambda_j)$ as described below.

$$\mathbf{J}(\lambda_j) = \begin{pmatrix} \mathbf{J}_1(\lambda_j) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_j) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_j}(\lambda_j) \end{pmatrix} \text{ with } \mathbf{J}_\star(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

- The largest Jordan block in $\mathbf{J}(\lambda_j)$ is $k_j \times k_j$, where $k_j = \mathit{index}(\lambda_j)$.
- The number of $i \times i$ Jordan blocks in $\mathbf{J}(\lambda_j)$ is given by

$$\nu_i(\lambda_j) = r_{i-1}(\lambda_j) - 2r_i(\lambda_j) + r_{i+1}(\lambda_j) \quad \text{with} \quad r_i(\lambda_j) = \mathit{rank}\left((\mathbf{A} - \lambda_j\mathbf{I})^i\right).$$

- Matrix $\mathbf{J}$ in (7.8.4) is called the ***Jordan form*** for $\mathbf{A}$. The *structure* of this form is unique in the sense that the number of Jordan segments in $\mathbf{J}$ as well as the number and sizes of the Jordan blocks in each segment is uniquely determined by the entries in $\mathbf{A}$. Furthermore, every matrix similar to $\mathbf{A}$ has the same Jordan structure—i.e., $\mathbf{A}, \mathbf{B} \in \mathcal{C}^{n\times n}$ are similar if and only if $\mathbf{A}$ and $\mathbf{B}$ have the same Jordan structure. The matrix $\mathbf{P}$ is not unique—see p. 594.

Example 7.8.1

Problem: Find the Jordan form for $\mathbf{A} = \begin{pmatrix} 5 & 4 & 0 & 0 & 4 & 3 \\ 2 & 3 & 1 & 0 & 5 & 1 \\ 0 & -1 & 2 & 0 & 2 & 0 \\ -8 & -8 & -1 & 2 & -12 & -7 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ -8 & -8 & -1 & 0 & -9 & -5 \end{pmatrix}$.

Solution: Computing the eigenvalues (which is the hardest part) reveals two distinct eigenvalues $\lambda_1 = 2$ and $\lambda_2 = -1$, so there are two Jordan segments in the Jordan form $\mathbf{J} = \begin{pmatrix} \mathbf{J}(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(-1) \end{pmatrix}$. Computing ranks $r_i(2) = rank\left((\mathbf{A} - 2\mathbf{I})^i\right)$ and $r_i(-1) = rank\left((\mathbf{A} + \mathbf{I})^i\right)$ until $r_k(\star) = r_{k+1}(\star)$ yields

$$\begin{aligned}
r_1(2) &= rank\,(\mathbf{A} - 2\mathbf{I}) &&= 4, & r_1(-1) &= rank\,(\mathbf{A} + \mathbf{I}) &&= 4,\\
r_2(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^2\right) &&= 3, & r_2(-1) &= rank\left((\mathbf{A} + \mathbf{I})^2\right) &&= 4,\\
r_3(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^3\right) &&= 2, & & & &\\
r_4(2) &= rank\left((\mathbf{A} - 2\mathbf{I})^4\right) &&= 2, & & & &
\end{aligned}$$

so $k_1 = index(\lambda_1) = 3$ and $k_2 = index(\lambda_2) = 1$. This tells us that the largest Jordan block in $\mathbf{J}(2)$ is 3×3, while the largest Jordan block in $\mathbf{J}(-1)$ is 1×1 so that $\mathbf{J}(-1)$ is a diagonal matrix (the associated eigenvalue is ***semisimple*** whenever this happens). Furthermore,

$$\begin{aligned}
\nu_3(2) &= r_2(2) - 2r_3(2) + r_4(2) = 1 &&\Longrightarrow \text{ one } 3 \times 3 \text{ block in } \mathbf{J}(2),\\
\nu_2(2) &= r_1(2) - 2r_2(2) + r_3(2) = 0 &&\Longrightarrow \text{ no } 2 \times 2 \text{ blocks in } \mathbf{J}(2),\\
\nu_1(2) &= r_0(2) - 2r_1(2) + r_2(2) = 1 &&\Longrightarrow \text{ one } 1 \times 1 \text{ block in } \mathbf{J}(2),\\
\nu_1(-1) &= r_0(-1) - 2r_1(-1) + r_2(-1) = 2 &&\Longrightarrow \text{ two } 1 \times 1 \text{ blocks in } \mathbf{J}(-1).
\end{aligned}$$

Therefore, $\mathbf{J}(2) = \left(\begin{array}{ccc|c} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ \hline 0 & 0 & 0 & 2 \end{array}\right)$ and $\mathbf{J}(-1) = \left(\begin{array}{c|c} -1 & 0 \\ \hline 0 & -1 \end{array}\right)$ so that

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}(2) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}(-1) \end{pmatrix} = \left(\begin{array}{ccc|c||cc} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 2 & 0 & 0 \\ \hline\hline 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right).$$

The above example suggests that determining the Jordan form for $\mathbf{A}_{n\times n}$ is straightforward, and perhaps even easy. In theory, it is—just find $\sigma(\mathbf{A})$, and calculate some ranks. But, in practice, both of these tasks can be difficult. To begin with, the rank of a matrix is a discontinuous function of its entries, and rank computed with floating-point arithmetic can vary with the algorithm used and is often different than rank computed with exact arithmetic (recall Exercise 2.2.4).

Furthermore, computing higher-index eigenvalues with floating-point arithmetic is fraught with peril. To see why, consider the matrix

$$\mathbf{L}(\epsilon) = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \epsilon & & & 0 \end{pmatrix}_{n\times n} \qquad \text{whose characteristic equation is} \qquad \lambda^n - \epsilon = 0.$$

For $\epsilon = 0$, zero is the only eigenvalue (and it has index n), but for all $\epsilon > 0$, there are n distinct eigenvalues given by $\epsilon^{1/n}e^{2k\pi i/n}$ for $k = 0, 1, \ldots, n-1$. For example, if $n = 32$, and if ϵ changes from 0 to 10^{-16}, then the eigenvalues of $\mathbf{L}(\epsilon)$ change in magnitude from 0 to $10^{-1/2} \approx .316$, which is substantial for such a small perturbation. Sensitivities of this kind present significant problems for floating-point algorithms. In addition to showing that high-index eigenvalues are sensitive to small perturbations, this example also shows that the Jordan structure is highly discontinuous. $\mathbf{L}(0)$ is in Jordan form, and there is just one Jordan block of size n, but for all $\epsilon \neq 0$, the Jordan form of $\mathbf{L}(\epsilon)$ is a diagonal matrix—i.e., there are n Jordan blocks of size 1×1. Lest you think that this example somehow is an isolated case, recall from Example 7.3.6 (p. 532) that *every* matrix in $\mathcal{C}^{n\times n}$ is arbitrarily close to a diagonalizable matrix.

All of the above observations make it clear that it's hard to have faith in a Jordan form that has been computed with floating-point arithmetic. Consequently, numerical computation of Jordan forms is generally avoided.

Example 7.8.2

The Jordan form of $\mathbf{A}$ conveys complete information about the eigenvalues of $\mathbf{A}$. For example, if the Jordan form for $\mathbf{A}$ is

$$\mathbf{J} = \left(\begin{array}{ccc|cc|cc|c|c} 4 & 1 & 0 & & & & & & \\ & 4 & 1 & & & & & & \\ & & 4 & & & & & & \\ \hline & & & 4 & 1 & & & & \\ & & & 0 & 4 & & & & \\ \hline & & & & & 3 & 1 & & \\ & & & & & 0 & 3 & & \\ \hline & & & & & & & 2 & \\ \hline & & & & & & & & 2 \end{array}\right),$$

then we know that

- ▷ $\mathbf{A}_{9\times 9}$ has three distinct eigenvalues, namely $\sigma(\mathbf{A}) = \{4, 3, 2\}$;
- ▷ $alg\ mult(4) = 5$, $alg\ mult(3) = 2$, and $alg\ mult(2) = 2$;
- ▷ $geo\ mult(4) = 2$, $geo\ mult(3) = 1$, and $geo\ mult(2) = 2$;

- ▷ $index(4) = 3$, $index(3) = 2$, and $index(2) = 1$;
- ▷ $\lambda = 2$ is a semisimple eigenvalue, so, while $\mathbf{A}$ is not diagonalizable, part of it is; i.e., the restriction $\mathbf{A}\big/_{N(\mathbf{A}-2\mathbf{I})}$ is a diagonalizable linear operator.

Of course, if both $\mathbf{P}$ and $\mathbf{J}$ are known, then $\mathbf{A}$ can be completely reconstructed from (7.8.4), but the point being made here is that only $\mathbf{J}$ is needed to reveal the eigenstructure along with the other similarity invariants of $\mathbf{A}$.

Now that the structure of the Jordan form $\mathbf{J}$ is known, the structure of the similarity transformation $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is easily revealed. Focus on a single $p \times p$ Jordan block $\mathbf{J}_\star(\lambda)$ contained in the Jordan segment $\mathbf{J}(\lambda)$ associated with an eigenvalue λ, and let $\mathbf{P}_\star = [\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,]$ be the portion of $\mathbf{P} = [\,\cdots\,|\,\mathbf{P}_\star\,|\,\cdots\,]$ that corresponds to the position of $\mathbf{J}_\star(\lambda)$ in $\mathbf{J}$. Notice that $\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}$ implies $\mathbf{A}\mathbf{P}_\star = \mathbf{P}_\star\mathbf{J}_\star(\lambda)$ or, equivalently,

$$\mathbf{A}[\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,] = [\,\mathbf{x}_1\,\mathbf{x}_2\,\cdots\,\mathbf{x}_p\,]\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}_{p\times p},$$

so equating columns on both sides of this equation produces

$$\begin{array}{lcll}
\mathbf{A}\mathbf{x}_1 = \lambda\mathbf{x}_1 & \Longrightarrow & \mathbf{x}_1 \text{ is an eigenvector} & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_1 = \mathbf{0}, \\
\mathbf{A}\mathbf{x}_2 = \mathbf{x}_1 + \lambda\mathbf{x}_2 & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_2 = \mathbf{x}_1 & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^2\,\mathbf{x}_2 = \mathbf{0}, \\
\mathbf{A}\mathbf{x}_3 = \mathbf{x}_2 + \lambda\mathbf{x}_3 & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_3 = \mathbf{x}_2 & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^3\,\mathbf{x}_3 = \mathbf{0}, \\
\quad\vdots & & \quad\vdots & \qquad\qquad\vdots \\
\mathbf{A}\mathbf{x}_p = \mathbf{x}_{p-1} + \lambda\mathbf{x}_p & \Longrightarrow & (\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_p = \mathbf{x}_{p-1} & \Longrightarrow \quad (\mathbf{A}-\lambda\mathbf{I})^p\,\mathbf{x}_p = \mathbf{0}.
\end{array}$$

In other words, the first column $\mathbf{x}_1$ in $\mathbf{P}_\star$ is a eigenvector for $\mathbf{A}$ associated with λ. We already knew there had to be exactly one independent eigenvector for each Jordan block because there are $t = \dim N\,(\mathbf{A}-\lambda\mathbf{I})$ Jordan blocks $\mathbf{J}_\star(\lambda)$, but now we know precisely where these eigenvectors are located in $\mathbf{P}$.

Vectors $\mathbf{x}$ such that $\mathbf{x} \in N\big((\mathbf{A}-\lambda\mathbf{I})^g\big)$ but $\mathbf{x} \notin N\big((\mathbf{A}-\lambda\mathbf{I})^{g-1}\big)$ are called ***generalized eigenvectors of order g*** associated with λ. So $\mathbf{P}_\star$ consists of an eigenvector followed by generalized eigenvectors of increasing order. Moreover, the columns of $\mathbf{P}_\star$ form a ***Jordan chain*** analogous to (7.7.2) on p. 576; i.e., $\mathbf{x}_i = (\mathbf{A}-\lambda\mathbf{I})^{p-i}\,\mathbf{x}_p$ implies $\mathbf{P}_\star$ must have the form

$$\mathbf{P}_\star = \big[\,(\mathbf{A}-\lambda\mathbf{I})^{p-1}\,\mathbf{x}_p\,|\,(\mathbf{A}-\lambda\mathbf{I})^{p-2}\,\mathbf{x}_p\,|\,\cdots\,|\,(\mathbf{A}-\lambda\mathbf{I})\,\mathbf{x}_p\,|\,\mathbf{x}_p\,\big]. \tag{7.8.5}$$

A complete set of Jordan chains associated with a given eigenvalue λ is determined in exactly the same way as Jordan chains for nilpotent matrices are

determined except that the nested subspaces $\mathcal{M}_i$ defined in (7.7.1) on p. 575 are redefined to be

$$\mathcal{M}_i = R\big((\mathbf{A}-\lambda\mathbf{I})^i\big) \cap N(\mathbf{A}-\lambda\mathbf{I}) \quad \text{for } i = 0, 1, \ldots, k, \tag{7.8.6}$$

where $k = \mathit{index}(\lambda)$. Just as in the case of nilpotent matrices, it follows that $\mathbf{0} = \mathcal{M}_k \subseteq \mathcal{M}_{k-1} \subseteq \cdots \subseteq \mathcal{M}_0 = N(\mathbf{A}-\lambda\mathbf{I})$ (see Exercise 7.8.8). Since $(\mathbf{A}-\lambda\mathbf{I})/_{\mathcal{N}((\mathbf{A}-\lambda\mathbf{I})^k)}$ is a nilpotent linear operator of index k (Example 5.10.4, p. 399), it can be argued that the same process used to build Jordan chains for nilpotent matrices can be used to build Jordan chains for a general eigenvalue λ. Below is a summary of the process adapted to the general case.

Constructing Jordan Chains

For each $\lambda \in \sigma(\mathbf{A}_{n\times n})$, set $\mathcal{M}_i = R\big((\mathbf{A}-\lambda\mathbf{I})^i\big) \cap N(\mathbf{A}-\lambda\mathbf{I})$ for $i = k-1, k-2, \ldots, 0$, where $k = \mathit{index}(\lambda)$.

- Construct a basis $\mathcal{B}$ for $N(\mathbf{A}-\lambda\mathbf{I})$.
  - ▷ Starting with any basis $\mathcal{S}_{k-1}$ for $\mathcal{M}_{k-1}$ (see p. 211), sequentially extend $\mathcal{S}_{k-1}$ with sets $\mathcal{S}_{k-2}$, $\mathcal{S}_{k-3}$, $\ldots$, $\mathcal{S}_0$ such that

$$\begin{array}{ll} \mathcal{S}_{k-1} & \text{is a basis for } \mathcal{M}_{k-1}, \\ \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} & \text{is a basis for } \mathcal{M}_{k-2}, \\ \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \mathcal{S}_{k-3} & \text{is a basis for } \mathcal{M}_{k-3}, \end{array}$$

    etc., until a basis $\mathcal{B} = \mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_0 = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t\}$ for $\mathcal{M}_0 = N(\mathbf{A}-\lambda\mathbf{I})$ is obtained (see Example 7.7.3 on p. 582).
- Build a Jordan chain on top of each eigenvector $\mathbf{b}_\star \in \mathcal{B}$.
  - ▷ For each eigenvector $\mathbf{b}_\star \in \mathcal{S}_i$, solve $(\mathbf{A}-\lambda\mathbf{I})^i\mathbf{x}_\star = \mathbf{b}_\star$ (a necessarily consistent system) for $\mathbf{x}_\star$, and construct a Jordan chain on top of $\mathbf{b}_\star$ by setting

$$\mathbf{P}_\star = \Big[\,(\mathbf{A}-\lambda\mathbf{I})^i\mathbf{x}_\star \,\Big|\, (\mathbf{A}-\lambda\mathbf{I})^{i-1}\mathbf{x}_\star \,\Big|\, \cdots \,\Big|\, (\mathbf{A}-\lambda\mathbf{I})\mathbf{x}_\star \,\Big|\, \mathbf{x}_\star\Big]_{(i+1)\times n}.$$

  - ▷ Each such $\mathbf{P}_\star$ corresponds to one Jordan block $\mathbf{J}_\star(\lambda)$ in the Jordan segment $\mathbf{J}(\lambda)$ associated with λ.
  - ▷ The first column in $\mathbf{P}_\star$ is an eigenvector, and subsequent columns are generalized eigenvectors of increasing order.
- If all such $\mathbf{P}_\star$'s for a given $\lambda_j \in \sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ are put in a matrix $\mathbf{P}_j$, and if $\mathbf{P} = \big[\,\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \cdots \,|\, \mathbf{P}_s\,\big]$, then $\mathbf{P}$ is a nonsingular matrix such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \operatorname{diag}\big(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s)\big)$ is in Jordan form as described on p. 590.

Example 7.8.3

Caution! Not every basis for $N(\mathbf{A}-\lambda\mathbf{I})$ can be used to build Jordan chains associated with an eigenvalue $\lambda \in \sigma(\mathbf{A})$. For example, the Jordan form of

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{pmatrix} \quad \text{is} \quad \mathbf{J} = \left(\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right)$$

because $\sigma(\mathbf{A}) = \{1\}$ and $\mathit{index}(1) = 2$. Consequently, if $\mathbf{P} = [\,\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3\,]$ is a nonsingular matrix such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$, then the derivation beginning on p. 593 leading to (7.8.5) shows that $\{\mathbf{x}_1, \mathbf{x}_2\}$ must be a Jordan chain such that $(\mathbf{A}-\mathbf{I})\mathbf{x}_1 = \mathbf{0}$ and $(\mathbf{A}-\mathbf{I})\mathbf{x}_2 = \mathbf{x}_1$, while $\mathbf{x}_3$ is another eigenvector (not dependent on $\mathbf{x}_1$). Suppose we try to build the Jordan chains in $\mathbf{P}$ by starting with the eigenvectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{7.8.7}$$

obtained by solving $(\mathbf{A}-\mathbf{I})\mathbf{x} = \mathbf{0}$ with straightforward Gauss–Jordan elimination. This naive approach fails because $\mathbf{x}_1 \notin R(\mathbf{A}-\mathbf{I})$ means $(\mathbf{A}-\mathbf{I})\mathbf{x}_2 = \mathbf{x}_1$ is an inconsistent system, so $\mathbf{x}_2$ cannot be determined. Similarly, $\mathbf{x}_3 \notin R(\mathbf{A}-\mathbf{I})$ insures that the same difficulty occurs if $\mathbf{x}_3$ is used in place of $\mathbf{x}_1$. In other words, even though the vectors in (7.8.7) constitute an otherwise perfectly good basis for $N(\mathbf{A}-\mathbf{I})$, they are not suitable for building Jordan chains. You are asked in Exercise 7.8.2 to find the correct basis for $N(\mathbf{A}-\mathbf{I})$ that will yield the Jordan chains that constitute $\mathbf{P}$.

Example 7.8.4

Problem: What do the results concerning the Jordan form for $\mathbf{A} \in \mathcal{C}^{n\times n}$ say about the decomposition of $\mathcal{C}^n$ into invariant subspaces?

Solution: Consider $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} = \operatorname{diag}\left(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s)\right)$, where the $\mathbf{J}(\lambda_j)$'s are the Jordan segments and $\mathbf{P} = \left[\,\mathbf{P}_1 \,|\, \mathbf{P}_2 \,|\, \cdots \,|\, \mathbf{P}_s\,\right]$ is a matrix of Jordan chains as described in (7.8.5) and on p. 594. If $\mathbf{A}$ is considered as a linear operator on $\mathcal{C}^n$, and if the set of columns in $\mathbf{P}_i$ is denoted by $\mathcal{J}_i$, then the results in §4.9 (p. 259) concerning invariant subspaces together with those in §5.9 (p. 383) about direct sum decompositions guarantee that each $R(\mathbf{P}_i)$ is an invariant subspace for $\mathbf{A}$ such that

$$\mathcal{C}^n = R(\mathbf{P}_1) \oplus R(\mathbf{P}_2) \oplus \cdots \oplus R(\mathbf{P}_s) \quad \text{and} \quad \mathbf{J}(\lambda_i) = \left[\mathbf{A}/_{R(\mathbf{P}_i)}\right]_{\mathcal{J}_i}.$$

More can be said. If $\mathit{alg\ mult}(\lambda_i) = m_i$ and $\mathit{index}(\lambda_i) = k_i$, then $\mathcal{J}_i$ is a linearly independent set containing m_i vectors, and the discussion surrounding

(7.8.5) insures that each column in $\mathcal{J}_i$ belongs to $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$. This coupled with the fact that $\dim N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)) = m_i$ (Exercise 7.8.7) implies that $\mathcal{J}_i$ is a basis for

$$R(\mathbf{P}_i) = N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big).$$

Consequently, each $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$ is an invariant subspace for $\mathbf{A}$ such that

$$\mathcal{C}^n = N\big((\mathbf{A}-\lambda_1\mathbf{I})^{k_1}\big) \oplus N\big((\mathbf{A}-\lambda_2\mathbf{I})^{k_2}\big) \oplus \cdots \oplus N\big((\mathbf{A}-\lambda_s\mathbf{I})^{k_s}\big)$$

and

$$\mathbf{J}(\lambda_i) = \left[\mathbf{A}\big/_{\mathcal{N}\left((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\right)}\right]_{\mathcal{J}_i}.$$

Of course, an even finer direct sum decomposition of $\mathcal{C}^n$ is possible because each Jordan segment is itself a block-diagonal matrix containing the individual Jordan blocks—the details are left to the interested reader.

Exercises for section 7.8

7.8.1. Find the Jordan form of the following matrix whose distinct eigenvalues are $\sigma(\mathbf{A}) = \{0, -1, 1\}$. Don't be frightened by the size of $\mathbf{A}$.

$$\mathbf{A} = \begin{pmatrix} -4 & -5 & -3 & 1 & -2 & 0 & 1 & -2 \\ 4 & 7 & 3 & -1 & 3 & 0 & -1 & 2 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 2 & -4 & 2 & 0 & -3 & 1 \\ -8 & -14 & -5 & 1 & -6 & 0 & 1 & -4 \\ 4 & 7 & 4 & -3 & 3 & -1 & -3 & 4 \\ 2 & -2 & -2 & 5 & -3 & 0 & 4 & -1 \\ 6 & 7 & 3 & 0 & 2 & 0 & 0 & 3 \end{pmatrix}.$$

7.8.2. For the matrix $\mathbf{A} = \begin{pmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{pmatrix}$ that was used in Example 7.8.3, use the technique described on p. 594 to construct a nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is in Jordan form.

7.8.3. Explain why $\mathit{index}(\lambda) \le \mathit{alg\ mult}(\lambda)$ for each $\lambda \in \sigma(\mathbf{A}_{n\times n})$.

7.8.4. Explain why $\mathit{index}(\lambda) = 1$ if and only if λ is a semisimple eigenvalue.

7.8.5. Prove that every square matrix is similar to its transpose. **Hint:** Consider the "reversal matrix" $\mathbf{R} = \begin{pmatrix} & & & 1 \\ & & 1 & \\ & \cdot^{\cdot^{\cdot}} & & \\ 1 & & & \end{pmatrix}$ obtained by reversing the order of the rows (or the columns) of the identity matrix $\mathbf{I}$.

7.8.6. Cayley–Hamilton Revisited. Prove the the Cayley–Hamilton theorem (pp. 509, 532) by means of the Jordan form; i.e., prove that every $\mathbf{A} \in \mathcal{C}^{n\times n}$ satisfies its own characteristic equation.

7.8.7. Prove that if λ is an eigenvalue of $\mathbf{A} \in \mathcal{C}^{n\times n}$ such that $index(\lambda) = k$ and $alg\ mult_{\mathbf{A}}(\lambda) = m$, then $\dim N\big((\mathbf{A} - \lambda\mathbf{I})^k\big) = m$. Is it also true that $\dim N\big((\mathbf{A} - \lambda\mathbf{I})^m\big) = m$?

7.8.8. Let λ_j be an eigenvalue of $\mathbf{A}$ with $index(\lambda_j) = k_j$. Prove that if $\mathcal{M}_i(\lambda_j) = R\big((\mathbf{A} - \lambda_j\mathbf{I})^i\big) \cap N(\mathbf{A} - \lambda_j\mathbf{I})$, then

$$\mathbf{0} = \mathcal{M}_{k_j}(\lambda_j) \subseteq \mathcal{M}_{k_{j-1}}(\lambda_j) \subseteq \cdots \subseteq \mathcal{M}_0(\lambda_j) = N(\mathbf{A} - \lambda_j\mathbf{I}).$$

7.8.9. Explain why $(\mathbf{A}-\lambda_j\mathbf{I})^i\mathbf{x} = \mathbf{b}(\lambda_j)$ must be a consistent system whenever $\lambda_j \in \sigma(\mathbf{A})$ and $\mathbf{b}(\lambda_j) \in \mathcal{S}_i(\lambda_j)$, where $\mathbf{b}(\lambda_j)$ and $\mathcal{S}_i(\lambda_j)$ are as defined on p. 594.

7.8.10. Does the result of Exercise 7.7.5 extend to nonnilpotent matrices? That is, if $\lambda \in \sigma(\mathbf{A})$ with $index(\lambda) = k > 1$, is $\mathcal{M}_{k-1} = R\big((\mathbf{A} - \lambda\mathbf{I})^{k-1}\big)$?

7.8.11. As defined in Exercise 5.8.15 (p. 380) and mentioned in Exercise 7.6.10 (p. 573), the ***Kronecker***[80] ***product*** (sometimes called *tensor product*,

[80] Leopold Kronecker (1823–1891) was born in Liegnitz, Prussia (now Legnica, Poland), to a wealthy business family that hired private tutors to educate him until he enrolled at Gymnasium at Liegnitz where his mathematical talents were recognized by Eduard Kummer (1810–1893), who became his mentor and lifelong colleague. Kronecker went to Berlin University in 1841 to earn his doctorate, writing on algebraic number theory, under the supervision of Dirichlet (p. 563). Rather than pursuing a standard academic career, Kronecker returned to Liegnitz to marry his cousin and become involved in his uncle's banking business. But he never lost his enjoyment of mathematics. After estate and business interests were left to others in 1855, Kronecker joined Kummer in Berlin who had just arrived to occupy the position vacated by Dirichlet's move to Göttingen. Kronecker didn't need a salary, so he didn't teach or hold a university appointment, but his research activities led to his election to the Berlin Academy in 1860. He declined the offer of the mathematics chair in Göttingen in 1868, but he eventually accepted the chair in Berlin that was vacated upon Kummer's retirement in 1883. Kronecker held the unconventional view that mathematics should be reduced to arguments that involve only integers and a finite number of steps, and he questioned the validity of nonconstructive existence proofs, so he didn't like the use of irrational or transcendental numbers. Kronecker became famous for saying that "God created the integers, all else is the work of man." Kronecker's significant influence led to animosity with people of differing philosophies such as Georg Cantor (1845–1918), whose publications Kronecker tried to block. Kronecker's small physical size was another sensitive issue. After Hermann Schwarz (p. 271), who was Kummer's son-in-law and a student of Weierstrass (p. 589), tried to make a joke involving Weierstrass's large physique by stating that "he who does not honor the Smaller, is not worthy of the Greater," Kronecker had no further dealings with Schwarz.

or *direct product*) of $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{p\times q}$ is the $mp \times nq$ matrix

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}.$$

(a) Assuming conformability, establish the following properties.

- $\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}$.
- $\mathbf{A}\otimes(\mathbf{B}+\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B})+(\mathbf{A}\otimes\mathbf{C})$.
- $(\mathbf{A}+\mathbf{B})\otimes\mathbf{C} = (\mathbf{A}\otimes\mathbf{C})+(\mathbf{B}\otimes\mathbf{C})$.
- $(\mathbf{A}_1\otimes\mathbf{B}_1)(\mathbf{A}_2\otimes\mathbf{B}_2)\cdots(\mathbf{A}_k\otimes\mathbf{B}_k) = (\mathbf{A}_1\cdots\mathbf{A}_k)\otimes(\mathbf{B}_1\cdots\mathbf{B}_k)$.
- $(\mathbf{A}\otimes\mathbf{B})^* = \mathbf{A}^*+\mathbf{B}^*$.
- $rank\,(\mathbf{A}\otimes\mathbf{B}) = (rank\,(\mathbf{A}))(rank\,(\mathbf{B}))$.

Assume $\mathbf{A}$ is $m\times m$ and $\mathbf{B}$ is $n\times n$ for the following.

- $trace\,(\mathbf{A}\otimes\mathbf{B}) = (trace\,(\mathbf{A}))(trace\,(\mathbf{B}))$.
- $(\mathbf{A}\otimes\mathbf{I}_n)(\mathbf{I}_m\otimes\mathbf{B}) = \mathbf{A}\otimes\mathbf{B} = (\mathbf{I}_m\otimes\mathbf{B})(\mathbf{A}\otimes\mathbf{I}_n)$.
- $\det(\mathbf{A}\otimes\mathbf{B}) = (\det(\mathbf{A}))^m(\det(\mathbf{B}))^n$.
- $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$.

(b) Let the eigenvalues of $\mathbf{A}_{m\times m}$ be denoted by λ_i and let the eigenvalues of $\mathbf{B}_{n\times n}$ be denoted by μ_j. Prove the following.

- The eigenvalues of $\mathbf{A}\otimes\mathbf{B}$ are the mn numbers $\{\lambda_i\mu_j\}_{i=1}^{m}{}_{j=1}^{n}$.
- The eigenvalues of $(\mathbf{A}\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{B})$ are $\{\lambda_i+\mu_j\}_{i=1}^{m}{}_{j=1}^{n}$.

7.8.12. Use part (b) of Exercise 7.8.11 along with the result of Exercise 7.6.10 (p. 573) to construct an alternate derivation of (7.6.8) on p. 566. That is, show that the n^2 eigenvalues of the discrete Laplacian $\mathbf{L}_{n^2\times n^2}$ described in Example 7.6.2 (p. 563) are given by

$$\lambda_{ij} = 4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right)+\sin^2\left(\frac{j\pi}{2(n+1)}\right)\right], \quad i,j = 1,2,\ldots,n.$$

Hint: Recall Exercise 7.2.18 (p. 522).

7.8.13. Determine the eigenvalues of the three-dimensional discrete Laplacian by using the formula from Exercise 7.6.10 (p. 573) that states

$$\mathbf{L}_{n^3\times n^3} = (\mathbf{I}_n\otimes\mathbf{I}_n\otimes\mathbf{A}_n)+(\mathbf{I}_n\otimes\mathbf{A}_n\otimes\mathbf{I}_n)+(\mathbf{A}_n\otimes\mathbf{I}_n\otimes\mathbf{I}_n).$$

7.9 FUNCTIONS OF NONDIAGONALIZABLE MATRICES

The development of functions of nondiagonalizable matrices parallels the development for functions of diagonal matrices that was presented in §7.3 except that the Jordan form is used in place of the diagonal matrix of eigenvalues. Recall from the discussion surrounding (7.3.5) on p. 526 that if $\mathbf{A} \in \mathcal{C}^{n\times n}$ is diagonalizable, say $\mathbf{A} = \mathbf{PDP}^{-1}$, where $\mathbf{D} = \operatorname{diag}\,(\lambda_1\mathbf{I}, \lambda_2\mathbf{I}, \ldots, \lambda_s\mathbf{I})$, and if $f(\lambda_i)$ exists for each λ_i, then $f(\mathbf{A})$ is defined to be

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} f(\lambda_1)\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & f(\lambda_2)\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & f(\lambda_s)\mathbf{I} \end{pmatrix}\mathbf{P}^{-1}.$$

The Jordan decomposition $\mathbf{A} = \mathbf{PJP}^{-1}$ described on p. 590 easily provides a generalization of this idea to nondiagonalizable matrices. If $\mathbf{J}$ is the Jordan form for $\mathbf{A}$, it's natural to define $f(\mathbf{A})$ by writing $f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1}$. However, there are a couple of wrinkles that need to be ironed out before this notion actually makes sense. First, we have to specify what we mean by $f(\mathbf{J})$—this is not as clear as $f(\mathbf{D})$ is for diagonal matrices. And after this is taken care of we need to make sure that $\mathbf{P}f(\mathbf{J})\mathbf{P}^{-1}$ is a uniquely defined matrix. This also is not clear because, as mentioned on p. 590, the transforming matrix $\mathbf{P}$ is not unique—it would not be good if for a given $\mathbf{A}$ you used one $\mathbf{P}$, and I used another, and this resulted in your $f(\mathbf{A})$ being different than mine.

Let's first make sense of $f(\mathbf{J})$. Assume throughout that $\mathbf{A} = \mathbf{PJP}^{-1} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ and where $\mathbf{J} = \operatorname{diag}\,(\mathbf{J}(\lambda_1), \mathbf{J}(\lambda_2), \ldots, \mathbf{J}(\lambda_s))$ is the Jordan form for $\mathbf{A}$ in which each segment $\mathbf{J}(\lambda_j)$ is a block-diagonal matrix containing one or more Jordan blocks. That is,

$$\mathbf{J}(\lambda_j) = \begin{pmatrix} \mathbf{J}_1(\lambda_j) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2(\lambda_j) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{t_j}(\lambda_j) \end{pmatrix} \quad \text{with} \quad \mathbf{J}_\star(\lambda_j) = \begin{pmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

We want to define $f(\mathbf{J})$ to be

$$f(\mathbf{J}) = \begin{pmatrix} f\big(\mathbf{J}(\lambda_1)\big) & & \\ & \ddots & \\ & & f\big(\mathbf{J}(\lambda_s)\big) \end{pmatrix} \quad \text{with} \quad f\big(\mathbf{J}(\lambda_j)\big) = \begin{pmatrix} \ddots & & \\ & f\big(\mathbf{J}_\star(\lambda_j)\big) & \\ & & \ddots \end{pmatrix},$$

but doing so requires that we give meaning to $f\big(\mathbf{J}_\star(\lambda_j)\big)$. To keep the notation from getting out of hand, let $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & \\ & \ddots & \ddots \\ & & \lambda \end{pmatrix}$ denote a generic $k \times k$ Jordan

block, and let's develop a definition of $f(\mathbf{J}_\star)$. Suppose for a moment that $f(z)$ is a function from $\mathcal{C}$ into $\mathcal{C}$ that has a Taylor series expansion about λ. That is, for some $r > 0$,

$$f(z) = f(\lambda) + f'(\lambda)(z-\lambda) + \frac{f''(\lambda)}{2!}(z-\lambda)^2 + \frac{f'''(\lambda)}{3!}(z-\lambda)^3 + \cdots \quad \text{for} \quad |z-\lambda| < r.$$

The representation (7.3.7) on p. 527 suggests that $f(\mathbf{J}_\star)$ should be defined as

$$f(\mathbf{J}_\star) = f(\lambda)\mathbf{I} + f'(\lambda)(\mathbf{J}_\star - \lambda\mathbf{I}) + \frac{f''(\lambda)}{2!}(\mathbf{J}_\star - \lambda\mathbf{I})^2 + \frac{f'''(\lambda)}{3!}(\mathbf{J}_\star - \lambda\mathbf{I})^3 + \cdots.$$

But since $\mathbf{N} = \mathbf{J}_\star - \lambda\mathbf{I}$ is nilpotent of index k, this series is just the finite sum

$$f(\mathbf{J}_\star) = \sum_{i=0}^{k-1} \frac{f^{(i)}(\lambda)}{i!}\mathbf{N}^i, \tag{7.9.1}$$

and this means that only $f(\lambda)$, $f'(\lambda)$, $\ldots$, $f^{(k-1)}(\lambda)$ are required to exist. Also,

$$\mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \quad \mathbf{N}^2 = \begin{pmatrix} 0 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}, \ldots, \mathbf{N}^{k-1} = \begin{pmatrix} 0 & 0 & \cdots & 1 \\ & \ddots & & \vdots \\ & & \ddots & 0 \\ & & & 0 \end{pmatrix},$$

so the representation of $f(\mathbf{J}_\star)$ in (7.9.1) can be elegantly expressed as follows.

Functions of Jordan Blocks

For a $k \times k$ Jordan block $\mathbf{J}_\star$ with eigenvalue λ, and for a function $f(z)$ such that $f(\lambda)$, $f'(\lambda)$, $\ldots$, $f^{(k-1)}(\lambda)$ exist, $f(\mathbf{J}_\star)$ is defined to be

$$f(\mathbf{J}_\star) = f\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix} = \begin{pmatrix} f(\lambda) & f'(\lambda) & \dfrac{f''(\lambda)}{2!} & \cdots & \dfrac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & f'(\lambda) & \ddots & \vdots \\ & & \ddots & \ddots & \dfrac{f''(\lambda)}{2!} \\ & & & f(\lambda) & f'(\lambda) \\ & & & & f(\lambda) \end{pmatrix}. \tag{7.9.2}$$

Every Jordan form $\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix}$ is a block-diagonal matrix composed of various Jordan blocks $\mathbf{J}_\star$, so (7.9.2) allows us to define $f(\mathbf{J}) = \begin{pmatrix} \ddots & & \\ & f(\mathbf{J}_\star) & \\ & & \ddots \end{pmatrix}$ as long as we pay attention to the fact that a sufficient number of derivatives of f are required to exist at the various eigenvalues. More precisely, if the size of the largest Jordan block associated with an eigenvalue λ is k (i.e., if $index(\lambda) = k$), then $f(\lambda), f'(\lambda), \ldots, f^{(k-1)}(\lambda)$ must exist in order for $f(\mathbf{J})$ to make sense.

Matrix Functions

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, let $k_i = index(\lambda_i)$.

- A function $f : \mathcal{C} \to \mathcal{C}$ is said to be defined (or to exist) at $\mathbf{A}$ when $f(\lambda_i), f'(\lambda_i), \ldots, f^{(k_i-1)}(\lambda_i)$ exist for each $\lambda_i \in \sigma(\mathbf{A})$.

- Suppose that $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, where $\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix}$ is in Jordan form with the $\mathbf{J}_\star$'s representing the various Jordan blocks described on p. 590. If f exists at $\mathbf{A}$, then the value of f at $\mathbf{A}$ is defined to be

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & f(\mathbf{J}_\star) & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1}, \tag{7.9.3}$$

where the $f(\mathbf{J}_\star)$'s are as defined in (7.9.2).

We still need to explain why (7.9.3) produces a uniquely defined matrix. The following argument will not only accomplish this purpose, but it will also establish an alternate expression for $f(\mathbf{A})$ that involves neither the Jordan form $\mathbf{J}$ nor the transforming matrix $\mathbf{P}$. Begin by partitioning $\mathbf{J}$ into its s Jordan segments as described on p. 590, and partition $\mathbf{P}$ and $\mathbf{P}^{-1}$ conformably as

$$\mathbf{P} = \Big(\mathbf{P}_1 \,|\, \cdots \,|\, \mathbf{P}_s\Big), \quad \mathbf{J} = \begin{pmatrix} \mathbf{J}(\lambda_1) & & \\ & \ddots & \\ & & \mathbf{J}(\lambda_s) \end{pmatrix}, \quad \text{and} \quad \mathbf{P}^{-1} = \begin{pmatrix} \mathbf{Q}_1 \\ \hline \vdots \\ \hline \mathbf{Q}_s \end{pmatrix}.$$

Define $\mathbf{G}_i = \mathbf{P}_i\mathbf{Q}_i$, and observe that if $k_i = index(\lambda_i)$, then $\mathbf{G}_i$ is the projector onto $N\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$. To see this, notice that $\mathbf{L}_i = \mathbf{J}(\lambda_i) - \lambda_i\mathbf{I}$ is nilpotent of index k_i, but $\mathbf{J}(\lambda_j) - \lambda_i\mathbf{I}$ is nonsingular when

$i \neq j$, so

$$(\mathbf{A} - \lambda_i \mathbf{I}) = \mathbf{P}(\mathbf{J} - \lambda_i \mathbf{I})\mathbf{P}^{-1} = \mathbf{P} \begin{pmatrix} \mathbf{J}(\lambda_1) - \lambda_i \mathbf{I} & & & & \\ & \ddots & & & \\ & & \mathbf{L}_i & & \\ & & & \ddots & \\ & & & & \mathbf{J}(\lambda_s) - \lambda_i \mathbf{I} \end{pmatrix} \mathbf{P}^{-1} \tag{7.9.4}$$

is a core-nilpotent decomposition as described on p. 397 (reordering the eigenvalues can put the nilpotent block $\mathbf{L}_i$ on the bottom to realize the form in (5.10.5)). Consequently, the results in Example 5.10.3 (p. 398) insure that $\mathbf{P}_i\mathbf{Q}_i = \mathbf{G}_i$ is the projector onto $N\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\big)$, and this is true for all similarity transformations that reduce $\mathbf{A}$ to $\mathbf{J}$. If $\mathbf{A}$ happens to be diagonalizable, then $k_i = 1$ for each i, and the matrices $\mathbf{G}_i = \mathbf{P}_i\mathbf{Q}_i$ are precisely the spectral projectors defined on p. 517. For this reason, there is no ambiguity in continuing to use the $\mathbf{G}_i$ notation, and we will continue to refer to the $\mathbf{G}_i$'s as ***spectral projectors.*** In the diagonalizable case, $\mathbf{G}_i$ projects onto the eigenspace associated with λ_i, and in the nondiagonalizable case $\mathbf{G}_i$ projects onto the generalized eigenspace associated with λ_i.

Now consider

$$f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1} = \mathbf{P} \begin{pmatrix} f\big(\mathbf{J}(\lambda_1)\big) & & \\ & \ddots & \\ & & f\big(\mathbf{J}(\lambda_s)\big) \end{pmatrix} \mathbf{P}^{-1} = \sum_{i=1}^{s} \mathbf{P}_i f\big(\mathbf{J}(\lambda_i)\big)\mathbf{Q}_i. \tag{7.9.5}$$

Since $f\big(\mathbf{J}(\lambda_i)\big) = \begin{pmatrix} \ddots & & \\ & f\big(\mathbf{J}_\star(\lambda_i)\big) & \\ & & \ddots \end{pmatrix}$, where the $\mathbf{J}_\star(\lambda_i)$'s are the Jordan blocks associated with λ_i, (7.9.2) insures that if $k_i = \mathit{index}(\lambda_i)$, then

$$f\big(\mathbf{J}(\lambda_i)\big) = f(\lambda_i)\mathbf{I} + f'(\lambda_i)\mathbf{L}_i + \frac{f''(\lambda_i)}{2!}\mathbf{L}_i^2 + \cdots + \frac{f^{(k_i-1)}(\lambda_i)}{(k_i-1)!}\mathbf{L}_i^{k_i-1},$$

where $\mathbf{L}_i = \mathbf{J}(\lambda_i) - \lambda_i\mathbf{I}$, and thus (7.9.5) becomes

$$f(\mathbf{A}) = \sum_{i=1}^{s} \mathbf{P}_i f\big(\mathbf{J}(\lambda_i)\big)\mathbf{Q}_i = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1} \frac{f^{(j)}(\lambda_i)}{j!}\mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i. \tag{7.9.6}$$

The terms $\mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i$ can be simplified by noticing that

$$\mathbf{P}^{-1}\mathbf{P} = \mathbf{I} \implies \mathbf{Q}_i\mathbf{P}_j = \begin{cases} \mathbf{I} & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \implies \mathbf{P}^{-1}\mathbf{G}_i = \begin{pmatrix} \mathbf{Q}_1 \\ \vdots \\ \mathbf{Q}_i \\ \vdots \\ \mathbf{Q}_s \end{pmatrix} \mathbf{P}_i\mathbf{Q}_i = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{Q}_i \\ \vdots \\ \mathbf{0} \end{pmatrix},$$

and by using this with (7.9.4) to conclude that

$$(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i = \mathbf{P}\begin{pmatrix} \big(\mathbf{J}(\lambda_1)-\lambda_i\mathbf{I}\big)^j & & & & \\ & \ddots & & & \\ & & \mathbf{L}_i^j & & \\ & & & \ddots & \\ & & & & \big(\mathbf{J}(\lambda_s)-\lambda_i\mathbf{I}\big)^j \end{pmatrix}\mathbf{P}^{-1}\mathbf{G}_i = \mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i. \quad (7.9.7)$$

Thus (7.9.6) can be written as

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i, \quad (7.9.8)$$

and this expression is independent of which similarity is used to reduce $\mathbf{A}$ to $\mathbf{J}$. Not only does (7.9.8) prove that $f(\mathbf{A})$ is uniquely defined, but it also provides a generalization of the spectral theorems for diagonalizable matrices given on pp. 517 and 526 because if $\mathbf{A}$ is diagonalizable, then each $k_i = 1$ so that (7.9.8) reduces to (7.3.6) on p. 526. Below is a formal summary along with some related properties.

Spectral Resolution of $f(\mathbf{A})$

For $\mathbf{A}\in\mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1,\lambda_2,\ldots,\lambda_s\}$ such that $k_i = \mathit{index}(\lambda_i)$, and for a function $f:\mathcal{C}\to\mathcal{C}$ such that $f(\lambda_i)$, $f'(\lambda_i)$, $\ldots$, $f^{(k_i-1)}(\lambda_i)$ exist for each $\lambda_i\in\sigma(\mathbf{A})$, the value of $f(\mathbf{A})$ is

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i, \quad (7.9.9)$$

where the ***spectral projectors*** $\mathbf{G}_i$'s have the following properties.

- $\mathbf{G}_i$ is the projector onto the generalized eigenspace $N\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$ along $R\big((\mathbf{A}-\lambda_i\mathbf{I})^{k_i}\big)$.
- $\mathbf{G}_1+\mathbf{G}_2+\cdots+\mathbf{G}_s = \mathbf{I}$. (7.9.10)
- $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ when $i\neq j$. (7.9.11)
- $\mathbf{N}_i = (\mathbf{A}-\lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{G}_i(\mathbf{A}-\lambda_i\mathbf{I})$ is nilpotent of index k_i. (7.9.12)
- If $\mathbf{A}$ is diagonalizable, then (7.9.9) reduces to (7.3.6) on p. 526, and the spectral projectors reduce to those described on p. 517.

Proof of (7.9.10)–(7.9.12). Property (7.9.10) results from using (7.9.9) with the function $f(z) = 1$, and property (7.9.11) is a consequence of

$$\mathbf{I} = \mathbf{P}^{-1}\mathbf{P} \implies \mathbf{Q}_i\mathbf{P}_j = \begin{cases} \mathbf{I} & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases} \tag{7.9.13}$$

To prove (7.9.12), establish that $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I})$ by noting that (7.9.13) implies $\mathbf{P}^{-1}\mathbf{G}_i = \big(\mathbf{0} \cdots \mathbf{Q}_i \cdots \mathbf{0}\big)^T$ and $\mathbf{G}_i\mathbf{P} = \big(\mathbf{0} \cdots \mathbf{P}_i \cdots \mathbf{0}\big)$. Use this with (7.9.4) to observe that $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{G}_i = \mathbf{P}_i\mathbf{L}_i\mathbf{Q}_i = \mathbf{G}_i(\mathbf{A} - \lambda_i\mathbf{I})$. Now

$$\mathbf{N}_i^j = (\mathbf{P}_i\mathbf{L}_i\mathbf{Q}_i)^j = \mathbf{P}_i\mathbf{L}_i^j\mathbf{Q}_i \quad \text{for } j = 1, 2, 3, \ldots,$$

and thus $\mathbf{N}_i$ is nilpotent of index k_i because $\mathbf{L}_i$ is nilpotent of index k_i. ∎

Example 7.9.1

A coordinate-free version of the representation in (7.9.3) results by separating the first-order terms in (7.9.9) from the higher-order terms to write

$$f(\mathbf{A}) = \sum_{i=1}^{s} \left[f(\lambda_i)\mathbf{G}_i + \sum_{j=1}^{k_i-1} \frac{f^{(j)}(\lambda_i)}{j!}\mathbf{N}_i^j \right].$$

Using the identity function $f(z) = z$ produces a coordinate-free version of the Jordan decomposition of $\mathbf{A}$ in the form

$$\mathbf{A} = \sum_{i=1}^{s} \big[\lambda_i\mathbf{G}_i + \mathbf{N}_i\big],$$

and this is the extension of (7.2.7) on p. 517 to the nondiagonalizable case. Another version of (7.9.9) results from lumping things into one matrix to write

$$f(\mathbf{A}) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1} f^{(j)}(\lambda_i)\mathbf{Z}_{ij}, \quad \text{where} \quad \mathbf{Z}_{ij} = \frac{(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i}{j!}. \tag{7.9.14}$$

The $\mathbf{Z}_{ij}$'s are often called the *component matrices* or the *constituent matrices.*

Example 7.9.2

Problem: Describe $f(\mathbf{A})$ for functions f defined at $\mathbf{A} = \begin{pmatrix} 6 & 2 & 8 \\ -2 & 2 & -2 \\ 0 & 0 & 2 \end{pmatrix}$.

Solution: $\mathbf{A}$ is block triangular, so it's easy to see that $\lambda_1 = 2$ and $\lambda_2 = 4$ are the two distinct eigenvalues with $\mathit{index}(\lambda_1) = 1$ and $\mathit{index}(\lambda_2) = 2$. Thus $f(\mathbf{A})$ exists for all functions such that $f(2)$, $f(4)$, and $f'(4)$ exist, in which case

$$f(\mathbf{A}) = f(2)\mathbf{G}_1 + f(4)\mathbf{G}_2 + f'(4)(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2.$$

The spectral projectors could be computed directly, but things are easier if some judicious choices of f are made. For example,

$$\left.\begin{array}{lll} f(z) = 1 & \Rightarrow & \mathbf{I} = f(\mathbf{A}) = \mathbf{G}_1 + \mathbf{G}_2 \\ f(z) = (z-4)^2 & \Rightarrow & (\mathbf{A} - 4\mathbf{I})^2 = f(\mathbf{A}) = 4\mathbf{G}_1 \end{array}\right\} \implies \begin{array}{l} \mathbf{G}_1 = (\mathbf{A} - 4\mathbf{I})^2/4, \\ \mathbf{G}_2 = \mathbf{I} - \mathbf{G}_1. \end{array}$$

Now that the spectral projectors are known, any function defined at $\mathbf{A}$ can be evaluated. For example, if $f(z) = z^{1/2}$, then

$$f(\mathbf{A}) = \sqrt{\mathbf{A}} = \sqrt{2}\mathbf{G}_1 + \sqrt{4}\mathbf{G}_2 + (1/2\sqrt{4})(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2 = \frac{1}{2}\begin{pmatrix} 5 & 1 & 7-2\sqrt{2} \\ -1 & 3 & 5-4\sqrt{2} \\ 0 & 0 & 2\sqrt{2} \end{pmatrix}.$$

This technique illustrated above is rather ad hoc, but it always works if a sufficient number of appropriate functions are used. For example, using $f(z) = z^p$ for $p = 0, 1, 2, \ldots$ will always produce a system of equations that will yield the component matrices $\mathbf{Z}_{ij}$ given in (7.9.14) because

$$\begin{aligned} &\text{for } f(z) = 1: && \mathbf{I} = \textstyle\sum \mathbf{Z}_{i0}, \\ &\text{for } f(z) = z: && \mathbf{A} = \textstyle\sum \lambda_i \mathbf{Z}_{i0} + \sum \mathbf{Z}_{i1}, \\ &\text{for } f(z) = z^2: && \mathbf{A}^2 = \textstyle\sum \lambda_i^2 \mathbf{Z}_{i0} + \sum 2\lambda_i \mathbf{Z}_{i1} + \sum 2\mathbf{Z}_{i2}, \\ &\qquad\vdots \end{aligned}$$

and this can be considered as a generalized Vandermonde linear system (p. 185)

$$\begin{pmatrix} 1 & \cdots & 1 & & & & & & \\ \lambda_1 & \cdots & \lambda_s & 1 & \cdots & 1 & & & \\ \lambda_1^2 & \cdots & \lambda_s^2 & 2\lambda_1 & \cdots & 2\lambda_s & 2 & \cdots & 2 \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ & \cdots & & & \cdots & & & \cdots & \cdots \end{pmatrix} \begin{pmatrix} \mathbf{Z}_{10} \\ \vdots \\ \mathbf{Z}_{s0} \\ \mathbf{Z}_{11} \\ \vdots \\ \mathbf{Z}_{s1} \\ \mathbf{Z}_{21} \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{A} \\ \mathbf{A}^2 \\ \mathbf{A}^3 \\ \vdots \end{pmatrix}$$

that can be solved for the $\mathbf{Z}_{ij}$'s. Other sets of polynomials such as

$$\{1,\ (z-\lambda_1)^{k_1},\ (z-\lambda_1)^{k_2}(z-\lambda_2)^{k_2},\ \ldots (z-\lambda_1)^{k_1}\cdots(z-\lambda_s)^{k_s}\}$$

will generate other linear systems that yield solutions containing the $\mathbf{Z}_{ij}$'s.

Example 7.9.3

Series Representations. Suppose that $\sum_{j=0}^{\infty} c_j(z-z_0)^j$ converges to $f(z)$ at each point inside a circle $|z - z_0| = r$, and suppose that $\mathbf{A}$ is a matrix such that $|\lambda_i - z_0| < r$ for each eigenvalue $\lambda_i \in \sigma(\mathbf{A})$.

Problem: Explain why $\sum_{j=0}^{\infty} c_j(\mathbf{A} - z_0\mathbf{I})^j$ converges to $f(\mathbf{A})$.

Solution: If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is in Jordan form as described on p. 601, then it's not difficult to argue that $\sum_{j=0}^{\infty} c_j(\mathbf{A} - z_0\mathbf{I})^j$ converges if and only if

$$\mathbf{P}^{-1}\Big(\sum_{j=0}^{\infty} c_j(\mathbf{A}-z_0\mathbf{I})^j\Big)\mathbf{P} = \sum_{j=0}^{\infty} c_j\mathbf{P}^{-1}(\mathbf{A}-z_0\mathbf{I})^j\mathbf{P} = \sum_{j=0}^{\infty} c_j(\mathbf{J}-z_0\mathbf{I})^j = \begin{pmatrix} \ddots & & \\ & \sum\limits_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j & \\ & & \ddots \end{pmatrix}$$

converges. Consequently, it suffices to prove that $\sum_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j$ converges to $f(\mathbf{J}_\star)$ for a generic $k \times k$ Jordan block

$$\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix} = \lambda\mathbf{I} + \mathbf{N}, \quad \text{where} \quad \mathbf{N} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & & 0 \end{pmatrix}_{k\times k}.$$

A standard theorem from analysis states that if $\sum_{j=0}^{\infty} c_j(z - z_0)^j$ converges to $f(z)$ when $|z - z_0| < r$, then the series may be differentiated term by term to yield series that converge to derivatives of f at points inside the circle of convergence. Consequently, for each $i = 0, 1, 2, \ldots,$

$$\frac{f^{(i)}(z)}{i!} = \sum_{j=0}^{\infty} c_j \binom{j}{i} (z - z_0)^{j-i} \quad \text{when} \quad |z - z_0| < r. \tag{7.9.15}$$

We know from (7.9.1) (with $f(z) = z^j$) that

$$(\mathbf{J}_\star - z_0\mathbf{I})^j = (\lambda - z_0)^j\mathbf{I} + \binom{j}{1}(\lambda - z_0)^{j-1}\mathbf{N} + \cdots + \binom{j}{k-1}(\lambda - z_0)^{j-(k-1)}\mathbf{N}^{k-1},$$

so this together with (7.9.15) produces

$$\begin{aligned}\sum_{j=0}^{\infty} c_j(\mathbf{J}_\star - z_0\mathbf{I})^j &= \left(\sum_{j=0}^{\infty} c_j(\lambda - z_0)^j\right)\mathbf{I} + \left(\sum_{j=0}^{\infty} c_j\binom{j}{1}(\lambda - z_0)^{j-1}\right)\mathbf{N} \\ &\qquad + \cdots + \left(\sum_{j=0}^{\infty} c_j\binom{j}{k-1}(\lambda - z_0)^{j-(k-1)}\right)\mathbf{N}^{k-1} \\ &= f(\lambda)\mathbf{I} + f'(\lambda)\mathbf{N} + \cdots + \frac{f^{(k-1)}}{(k-1)!}(\lambda)\mathbf{N}^{k-1} = f(\mathbf{J}_*).\end{aligned}$$

Note: The result of this example validates the statements made on p. 527.

Example 7.9.4

All Matrix Functions Are Polynomials. It was pointed out on p. 528 that if $\mathbf{A}$ is diagonalizable, and if $f(\mathbf{A})$ exists, then there is a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$, and you were asked in Exercise 7.3.7 (p. 539) to use the Cayley–Hamilton theorem (pp. 509, 532) to extend this property to nondiagonalizable matrices for functions that have an infinite series expansion. We can now see why this is true in general.

Problem: For a function f defined at $\mathbf{A} \in \mathcal{C}^{n\times n}$, exhibit a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$.

Solution: Suppose that $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$. The trick is to find a polynomial $p(z)$ such that for each $i = 1, 2, \ldots, s$,

$$p(\lambda_i) = f(\lambda_i), \quad p'(\lambda_i) = f'(\lambda_i), \quad \ldots, \quad p^{(k_i-1)}(\lambda_i) = f^{(k_i-1)}(\lambda_i) \tag{7.9.16}$$

because if such a polynomial exists, then (7.9.9) guarantees that

$$p(\mathbf{A}) = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{p^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i \mathbf{I})^j \mathbf{G}_i = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A} - \lambda_i \mathbf{I})^j \mathbf{G}_i = f(\mathbf{A}).$$

Since there are $k = \sum_{i=1}^{s} k_i$ equations in (7.9.16) to be satisfied, let's look for a polynomial of the form

$$p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_{k-1} z^{k-1}$$

by writing the equations in (7.9.16) as the following $k \times k$ linear system $\mathbf{Hx} = \mathbf{f}$:

$$\begin{array}{c}
p(\lambda_1) = f(\lambda_1) \\
\vdots \\
p(\lambda_s) = f(\lambda_s) \\
\hline
\vdots \\
p'(\lambda_i) = f'(\lambda_i) \\
\vdots \\
\hline
\vdots \\
p''(\lambda_i) = f''(\lambda_i) \\
\vdots \\
\hline
\vdots
\end{array}
\begin{array}{c}
\Rightarrow \\
\\
\Rightarrow \\
\\
\Rightarrow \\
\\
\Rightarrow \\
\\
\\
\end{array}
\left(\begin{array}{cccccc}
1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & \cdots & \lambda_1^{k-1} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
1 & \lambda_s & \lambda_s^2 & \lambda_s^3 & \cdots & \lambda_s^{k-1} \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & 1 & 2\lambda_i & 3\lambda_i^2 & \cdots & (k-1)\lambda_i^{k-2} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots \\
0 & 0 & 2 & 6\lambda_i & \cdots & (k-1)(k-2)\lambda_i^{(k-3)} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
\hline
\vdots & \vdots & \vdots & \vdots & & \vdots
\end{array}\right)
\left(\begin{array}{c}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\vdots \\
\vdots \\
\vdots \\
\alpha_{k-1}
\end{array}\right)
=
\left(\begin{array}{c}
f(\lambda_1) \\
\vdots \\
f(\lambda_s) \\
\hline
\vdots \\
f'(\lambda_i) \\
\vdots \\
\hline
\vdots \\
f''(\lambda_i) \\
\vdots \\
\hline
\vdots
\end{array}\right).$$

The coefficient matrix $\mathbf{H}$ can be proven to be nonsingular because the rows in each segment of $\mathbf{H}$ are linearly independent. The rows in the top segment of $\mathbf{H}$ are a subset of rows from a Vandermonde matrix (p. 185), while the nonzero portion of each succeeding segment has the form $\mathbf{VD}$, where the rows of $\mathbf{V}$ are a subset of rows from a Vandermonde matrix and $\mathbf{D}$ is a nonsingular diagonal matrix. Consequently, $\mathbf{Hx} = \mathbf{f}$ has a unique solution, and thus there is a unique polynomial $p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_{k-1} z^{k-1}$ that satisfies the conditions in (7.9.16). This polynomial $p(z)$ is called the ***Hermite interpolation polynomial***, and it has the property that $f(\mathbf{A}) = p(\mathbf{A})$.

Example 7.9.5

Functional Identities. Scalar functional identities generally extend to the matrix case. For example, the scalar identity $\sin^2 z + \cos^2 z = 1$ extends to matrices as $\sin^2 \mathbf{Z} + \cos^2 \mathbf{Z} = \mathbf{I}$, and this is valid for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$. While it's possible to prove such identities on a case-by-case basis by using (7.9.3) or (7.9.9), there is a more robust approach that is described below.

For two functions f_1 and f_2 from $\mathcal{C}$ into $\mathcal{C}$ and for a polynomial $p(x,y)$ in two variables, let h be the composition defined by $h(z) = p\big(f_1(z),\, f_2(z)\big)$. If $\mathbf{A}_{n\times n}$ has eigenvalues $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$, and if h is defined at $\mathbf{A}$, then we are allowed to assert that $h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$ because Example 7.9.4 insures that there are polynomials $g(z)$ and $q(z)$ such that $h(\mathbf{A}) = g(\mathbf{A})$ and $p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big) = q(\mathbf{A})$, where for each $\lambda_i \in \sigma(\mathbf{A})$,

$$g^{(j)}(\lambda_i) = h^{(j)}(\lambda_i) = \left.\frac{d^j\big[p\big(f_1(z),\, f_2(z)\big)\big]}{dz^j}\right|_{z=\lambda_i} = q^{(j)}(\lambda_i) \quad \text{for } j = 0, 1, \ldots, k_i - 1,$$

so $g(\mathbf{A}) = q(\mathbf{A})$, and thus $h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$. To build functional identities for $\mathbf{A}$, choose f_1 and f_2 in $h(z) = p\big(f_1(z),\, f_2(z)\big)$ that will make

$$h(\lambda_i) = h'(\lambda_i) = h''(\lambda_i) = \cdots = h^{(k_i-1)}(\lambda_i) = 0 \quad \text{for each} \quad \lambda_i \in \sigma(\mathbf{A}),$$

thereby insuring that $\mathbf{0} = \mathbf{h}(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big)$. This technique produces a plethora of functional identities. For example, using

$$\left\{\begin{array}{l} f_1(z) = \sin^2 z \\ f_2(z) = \cos^2 z \\ p(x,y) = x^2 + y^2 - 1 \end{array}\right\} \text{ produces } h(z) = p\big(f_1(z),\, f_2(z)\big) = \sin^2 z + \cos^2 z - 1.$$

Since $h(z) = 0$ for all $z \in \mathcal{C}$, it follows that $h(\mathbf{Z}) = \mathbf{0}$ for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$, and thus $\sin^2 \mathbf{Z} + \cos^2 \mathbf{Z} = \mathbf{I}$ for all $\mathbf{Z} \in \mathcal{C}^{n\times n}$. It's evident that this technique can be extended to include any number of functions $f_1, f_2, \ldots, f_m$ with a polynomial $p(x_1, x_2, \ldots, x_m)$ to produce even more complicated relationships.

Example 7.9.6

Systems of Differential Equations Revisited. The purpose here is to extend the discussion in §7.4 to cover the nondiagonalizable case. Write the system of differential equations in (7.4.1) on p. 541 in matrix form as

$$\mathbf{u}'(t) = \mathbf{A}_{n\times n}\mathbf{u}(t) \quad \text{with} \quad \mathbf{u}(0) = \mathbf{c}, \tag{7.9.17}$$

but this time don't assume that $\mathbf{A}_{n\times n}$ is diagonalizable—suppose instead that $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $index(\lambda_i) = k_i$. The development parallels that

for the diagonalizable case, but $e^{\mathbf{A}t}$ is now a little more complicated than (7.4.2). Using $f(z) = e^{zt}$ in (7.9.3) and (7.9.2) yields

$$e^{\mathbf{A}t} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & e^{\mathbf{J}_\star t} & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1} \text{ with } e^{\mathbf{J}_\star t} = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & \dfrac{t^2e^{\lambda t}}{2!} & \cdots & \dfrac{t^{k-1}e^{\lambda t}}{(k-1)!} \\ & e^{\lambda t} & te^{\lambda t} & \ddots & \vdots \\ & & \ddots & \ddots & \dfrac{t^2e^{\lambda t}}{2!} \\ & & & e^{\lambda t} & te^{\lambda t} \\ & & & & e^{\lambda t} \end{pmatrix}, \quad (7.9.18)$$

while setting $f(z) = e^{zt}$ in (7.9.9) produces

$$e^{\mathbf{A}t} = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{t^j e^{\lambda_i t}}{j!}(\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i. \quad (7.9.19)$$

Either of these can be used to show that the three properties (7.4.3)–(7.4.5) on p. 541 still hold. In particular, $d\,e^{\mathbf{A}t}/dt = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$, so, just as in the diagonalizable case, $\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{c}$ is the unique solution of (7.9.17) (the uniqueness argument given in §7.4 remains valid). In the diagonalizable case, the solution of (7.9.17) involves only the eigenvalues and eigenvectors of $\mathbf{A}$ as described in (7.4.7) on p. 542, but generalized eigenvectors are needed for the nondiagonalizable case. Using (7.9.19) yields the solution to (7.9.17) as

$$\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{c} = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{t^j e^{\lambda_i t}}{j!}\mathbf{v}_j(\lambda_i), \quad \text{where } \mathbf{v}_j(\lambda_i) = (\mathbf{A} - \lambda_i\mathbf{I})^j\mathbf{G}_i\mathbf{c}. \quad (7.9.20)$$

Each $\mathbf{v}_{k_i-1}(\lambda_i)$ is an eigenvector associated with λ_i because $(\mathbf{A} - \lambda_i\mathbf{I})^{k_i}\mathbf{G}_i = \mathbf{0}$, and $\{\mathbf{v}_{k_i-2}(\lambda_i), \ldots, \mathbf{v}_1(\lambda_i), \mathbf{v}_0(\lambda_i)\}$ is an associated chain of generalized eigenvectors. The behavior of the solution (7.9.20) as $t \to \infty$ is similar but not identical to that discussed on p. 544 because for $\lambda = x + iy$ and $t > 0$,

$$t^j e^{\lambda t} = t^j e^{xt}(\cos yt + i\sin yt) \to \begin{cases} 0 & \text{if } x < 0, \\ \text{unbounded} & \text{if } x \geq 0 \text{ and } j > 0, \\ \text{oscillates indefinitely} & \text{if } x = j = 0 \text{ and } y \neq 0, \\ 1 & \text{if } x = y = j = 0. \end{cases}$$

In particular, if $\operatorname{Re}(\lambda_i) < 0$ for every $\lambda_i \in \sigma(\mathbf{A})$, then $\mathbf{u}(t) \to \mathbf{0}$ for every initial vector $\mathbf{c}$, in which case the system is said to be ***stable***.

- **Nonhomogeneous Systems.** It can be verified by direct manipulation that the solution of $\mathbf{u}'(t) = \mathbf{A}\mathbf{u}(t) + \mathbf{f}(t)$ with $\mathbf{u}(t_0) = \mathbf{c}$ is given by

$$\mathbf{u}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{c} + \int_{t_0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{f}(\tau)d\tau.$$

Example 7.9.7

Nondiagonalizable Mixing Problem. To make the point that even simple problems in nature can be nondiagonalizable, consider three V gallon tanks as shown in Figure 7.9.1 that are initially full of polluted water in which the ith tank contains c_i lbs of a pollutant. In an attempt to flush the pollutant out, all spigots are opened at once allowing fresh water at the rate of r gal/sec to flow into the top of tank #3, while r gal/sec flow from its bottom into the top of tank #2, and so on.

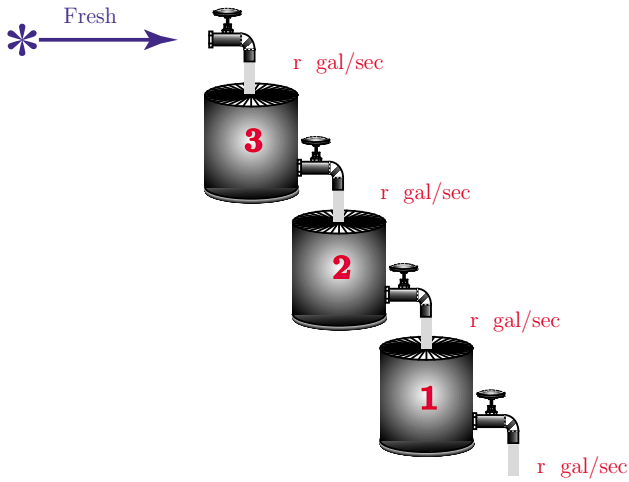


FIGURE 7.9.1

Problem: How many pounds of the pollutant are in each tank at any finite time $t > 0$ when instantaneous and continuous mixing occurs?

Solution: If $u_i(t)$ denotes the number of pounds of pollutant in tank i at time $t > 0$, then the concentration of pollutant in tank i at time t is $u_i(t)/V$ lbs/gal, so the model $u_i'(t) =$ (lbs/sec) coming in$-$(lbs/sec) going out produces the nondiagonalizable system:

$$\begin{pmatrix} u_1'(t) \\ u_2'(t) \\ u_3'(t) \end{pmatrix} = \frac{r}{V} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix}, \text{ or } \mathbf{u}' = \mathbf{A}\mathbf{u} \text{ with } \mathbf{u}(0) = \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}.$$

This setup is almost the same as that in Exercise 3.5.11 (p. 104). Notice that $\mathbf{A}$ is simply a scalar multiple of a single Jordan block $\mathbf{J}_\star = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}$, so $e^{\mathbf{A}t}$ is easily determined by replacing t by rt/V and λ by -1 in the second equation of (7.9.18) to produce

$$e^{\mathbf{A}t} = e^{(rt/V)\mathbf{J}_\star} = e^{-rt/V} \begin{pmatrix} 1 & rt/V & (rt/V)^2/2 \\ 0 & 1 & rt/V \\ 0 & 0 & 1 \end{pmatrix}.$$

Therefore,

$$\mathbf{u}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{c} = \mathrm{e}^{-rt/V}\begin{pmatrix} c_1 + c_2(rt/V) + c_3\,(rt/V)^2/2 \\ c_2 + c_3(rt/V) \\ c_3 \end{pmatrix},$$

and, just as common sense dictates, the pollutant is never completely flushed from the tanks in finite time. Only in the limit does each $u_i \to 0$, and it's clear that the rate at which $u_1 \to 0$ is slower than the rate at which $u_2 \to 0$, which in turn is slower than the rate at which $u_3 \to 0$.

Example 7.9.8

The Cauchy integral formula is an elegant result from complex analysis stating that if $f : \mathcal{C} \to \mathcal{C}$ is analytic in and on a simple closed contour $\Gamma \subset \mathcal{C}$ with positive (counterclockwise) orientation, and if ξ_0 is interior to Γ, then

$$f(\xi_0) = \frac{1}{2\pi\mathrm{i}}\int_\Gamma \frac{f(\xi)}{\xi - \xi_0}d\xi \quad \text{and} \quad f^{(j)}(\xi_0) = \frac{j!}{2\pi\mathrm{i}}\int_\Gamma \frac{f(\xi)}{(\xi-\xi_0)^{j+1}}d\xi. \tag{7.9.21}$$

These formulas produce analogous representations of matrix functions. Suppose that $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ and $index(\lambda_i) = k_i$. For a complex variable ξ, the ***resolvent of*** $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the matrix

$$\mathbf{R}(\xi) = (\xi\mathbf{I} - \mathbf{A})^{-1}.$$

If $\xi \notin \sigma(\mathbf{A})$, then $r(z) = (\xi - z)^{-1}$ is defined at $\mathbf{A}$ with $r(\mathbf{A}) = \mathbf{R}(\xi)$, so the spectral resolution theorem (p. 603) can be used to write

$$\mathbf{R}(\xi) = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{r^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i = \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{1}{(\xi-\lambda_i)^{j+1}}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i.$$

If $\sigma(\mathbf{A})$ is in the interior of a simple closed contour Γ, and if the contour integral of a matrix is defined by entrywise integration, then (7.9.21) produces

$$\begin{aligned}
\frac{1}{2\pi\mathrm{i}}\int_\Gamma f(\xi)(\xi\mathbf{I}-\mathbf{A})^{-1}d\xi &= \frac{1}{2\pi\mathrm{i}}\int_\Gamma f(\xi)\mathbf{R}(\xi)d\xi \\
&= \frac{1}{2\pi\mathrm{i}}\int_\Gamma \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f(\xi)}{(\xi-\lambda_i)^{j+1}}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i d\xi \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\left[\frac{1}{2\pi\mathrm{i}}\int_\Gamma \frac{f(\xi)}{(\xi-\lambda_i)^{j+1}}d\xi\right](\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{f^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i = f(\mathbf{A}).
\end{aligned}$$

- In other words, if Γ is a simple closed contour containing $\sigma(\mathbf{A})$ in its interior, then
$$f(\mathbf{A}) = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma} f(\xi)(\xi \mathbf{I} - \mathbf{A})^{-1} d\xi \tag{7.9.22}$$
whenever f is analytic in and on Γ. Since this formula makes sense for general linear operators, it is often adopted as a definition for $f(\mathbf{A})$ in more general settings.

- Furthermore, if Γ_i is a simple closed contour enclosing λ_i but excluding all other eigenvalues of $\mathbf{A}$, then the ith spectral projector is given by
$$\mathbf{G}_i = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} \mathbf{R}(\xi) d\xi = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi \qquad \text{(Exercise 7.9.19).}$$

Exercises for section 7.9

7.9.1. Lake $\#i$ in a closed system of three lakes of equal volume V initially contains c_i lbs of a pollutant. If the water in the system is circulated at rates (gal/sec) as indicated in Figure 7.9.2, find the amount of pollutant in each lake at time $t > 0$ (assume continuous mixing), and then determine the pollution in each lake in the long run.

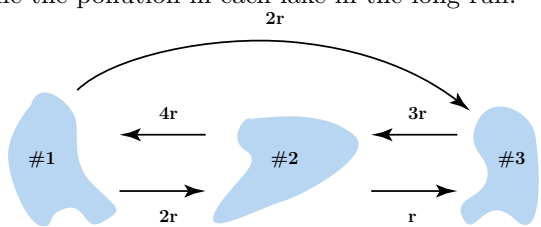


FIGURE 7.9.2

7.9.2. Suppose that $\mathbf{A} \in \mathcal{C}^{n \times n}$ has eigenvalues λ_i with $index(\lambda_i) = k_i$. Explain why the ith spectral projector is given by
$$\mathbf{G}_i = f_i(\mathbf{A}), \quad \text{where} \quad f_i(z) = \begin{cases} 1 & \text{when } z = \lambda_i, \\ 0 & \text{otherwise.} \end{cases}$$

7.9.3. Explain why each spectral projector $\mathbf{G}_i$ can be expressed as a polynomial in $\mathbf{A}$.

7.9.4. If $\sigma(\mathbf{A}_{n \times n}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$ with $k_i = index(\lambda_i)$, explain why
$$\mathbf{A}^k = \sum_{i=1}^{s} \sum_{j=0}^{k_i - 1} \binom{k}{j} \lambda_i^{k-j} (\mathbf{A} - \lambda_i \mathbf{I})^j \mathbf{G}_i.$$

7.9.5. With the convention that $\binom{k}{j} = 0$ for $j > k$, explain why

$$\begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix}^k_{m\times m} = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \lambda^k & \binom{k}{1}\lambda^{k-1} & \ddots & \vdots \\ & & \ddots & \ddots & \binom{k}{2}\lambda^{k-2} \\ & & & \lambda^k & \binom{k}{1}\lambda^{k-1} \\ & & & & \lambda^k \end{pmatrix}.$$

7.9.6. Determine $e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 6 & 2 & 8 \\ -2 & 2 & -2 \\ 0 & 0 & 2 \end{pmatrix}$.

7.9.7. For $f(z) = 4\sqrt{z} - 1$, determine $f(\mathbf{A})$ when $\mathbf{A} = \begin{pmatrix} -3 & -8 & -9 \\ 5 & 11 & 9 \\ -1 & -2 & 1 \end{pmatrix}$.

7.9.8. (a) Explain why every nonsingular $\mathbf{A} \in \mathcal{C}^{n\times n}$ has a square root.

(b) Give necessary and sufficient conditions for the existence of $\sqrt{\mathbf{A}}$ when $\mathbf{A}$ is singular.

7.9.9. **Spectral Mapping Property.** Prove that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(f(\lambda), \mathbf{x})$ is an eigenpair for $f(\mathbf{A})$ whenever $f(\mathbf{A})$ exists. Does it also follow that $alg\ mult_{\mathbf{A}}(\lambda) = alg\ mult_{f(\mathbf{A})}(f(\lambda))$?

7.9.10. Let f be defined at $\mathbf{A}$, and let $\lambda \in \sigma(\mathbf{A})$. Give an example or an explanation of why the following statements are *not* necessarily true.

(a) $f(\mathbf{A})$ is similar to $\mathbf{A}$.

(b) $geo\ mult_{\mathbf{A}}(\lambda) = geo\ mult_{f(\mathbf{A})}(f(\lambda))$.

(c) $index_{\mathbf{A}}(\lambda) = index_{f(\mathbf{A})}(f(\lambda))$.

7.9.11. Explain why $\mathbf{A}f(\mathbf{A}) = f(\mathbf{A})\mathbf{A}$ whenever $f(\mathbf{A})$ exists.

7.9.12. Explain why a function f is defined at $\mathbf{A} \in \mathcal{C}^{n\times n}$ if and only if f is defined at $\mathbf{A}^T$, and then prove that $f(\mathbf{A}^T) = \left[f(\mathbf{A})\right]^T$. Why can't $(\star)^*$ be used in place of $(\star)^T$?

7.9.13. Use the technique of Example 7.9.5 (p. 608) to establish the following identities.

(a) $e^{\mathbf{A}}e^{-\mathbf{A}} = \mathbf{I}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$.

(b) $e^{\alpha\mathbf{A}} = \left(e^{\mathbf{A}}\right)^{\alpha}$ for all $\alpha \in \mathcal{C}$ and $\mathbf{A} \in \mathcal{C}^{n\times n}$.

(c) $e^{i\mathbf{A}} = \cos\mathbf{A} + i\sin\mathbf{A}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$.

7.9.14. (a) Show that if $\mathbf{AB} = \mathbf{BA}$, then $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$.

(b) Give an example to show that $e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}}e^{\mathbf{B}}$ in general.

7.9.15. Find the Hermite interpolation polynomial $p(z)$ as described in Example 7.9.4 such that $p(\mathbf{A}) = e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ -3 & -2 & -1 \\ -3 & -2 & -1 \end{pmatrix}$.

7.9.16. The Cayley–Hamilton theorem (pp. 509, 532) says that every $\mathbf{A} \in \mathcal{C}^{n\times n}$ satisfies its own characteristic equation, and this guarantees that $\mathbf{A}^{n+j}$ $(j = 0, 1, 2, \ldots)$ can be expressed as a polynomial in $\mathbf{A}$ of at most degree $n-1$. Since $f(\mathbf{A})$ is always a polynomial in $\mathbf{A}$, the Cayley–Hamilton theorem insures that $f(\mathbf{A})$ can be expressed as a polynomial in $\mathbf{A}$ of at most degree $n-1$. Such a polynomial can be determined whenever $f^{(j)}(\lambda_i)$, $j = 0, 1, \ldots, a_i - 1$ exists for each $\lambda_i \in \sigma(\mathbf{A})$, where $a_i = \mathit{alg\ mult}(\lambda_i)$. The strategy is the same as that in Example 7.9.4 except that a_i is used in place of k_i. If we can find a polynomial $p(z) = \alpha_0 + \alpha_1 z + \cdots + \alpha_{n-1}z^{n-1}$ such that for each $\lambda_i \in \sigma(\mathbf{A})$,

$$p(\lambda_i) = f(\lambda_i), \quad p'(\lambda_i) = f'(\lambda_i), \quad \ldots, \quad p^{(a_i-1)}(\lambda_i) = f^{(a_i-1)}(\lambda_i),$$

then $p(\mathbf{A}) = f(\mathbf{A})$. Why? These equations are an $n \times n$ linear system with the α_i's as the unknowns, and, for the same reason outlined in Example 7.9.4, a solution is always possible.

(a) What advantages and disadvantages does this approach have with respect to the approach in Example 7.9.4?

(b) Use this method to find a polynomial $p(z)$ such that $p(\mathbf{A}) = e^{\mathbf{A}}$ for $\mathbf{A} = \begin{pmatrix} 3 & 2 & 1 \\ -3 & -2 & -1 \\ -3 & -2 & -1 \end{pmatrix}$. Compare with Exercise 7.9.15.

7.9.17. Show that if f is a function defined at

$$\mathbf{A} = \begin{pmatrix} \alpha & \beta & \gamma \\ 0 & \alpha & \beta \\ 0 & 0 & \alpha \end{pmatrix} = \alpha\mathbf{I} + \beta\mathbf{N} + \gamma\mathbf{N}^2, \quad \text{where} \quad \mathbf{N} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

then $f(\mathbf{A}) = f(\alpha)\mathbf{I} + \beta f'(\alpha)\mathbf{N} + \left[\gamma f'(\alpha) + \dfrac{\beta^2 f''(\alpha)}{2!}\right]\mathbf{N}^2$.

7.9.18. Composition of Matrix Functions. If $h(z) = f(g(z))$, where f and g are functions such that $g(\mathbf{A})$ and $f\big(g(\mathbf{A})\big)$ each exist, then $h(\mathbf{A}) = f\big(g(\mathbf{A})\big)$. However, it's not legal to prove this simply by saying "replace z by $\mathbf{A}$." One way to prove that $h(\mathbf{A}) = f\big(g(\mathbf{A})\big)$ is to demonstrate that $h(\mathbf{J}_\star) = f\big(g(\mathbf{J}_\star)\big)$ for a generic Jordan block and then invoke (7.9.3). Do this for a 3×3 Jordan block—the generalization to $k \times k$ blocks is similar. That is, let $h(z) = f(g(z))$, and use Exercise 7.9.17 to prove that if $g(\mathbf{J}_\star)$ and $f\big(g(\mathbf{J}_\star)\big)$ each exist, then

$$h(\mathbf{J}_\star) = f\big(g(\mathbf{J}_\star)\big) \quad \text{for} \quad \mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}.$$

7.9.19. Prove that if Γ_i is a simple closed contour enclosing $\lambda_i \in \sigma(\mathbf{A})$ but excluding all other eigenvalues of $\mathbf{A}$, then the ith spectral projector is

$$\mathbf{G}_i = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi = \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_i} \mathbf{R}(\xi) d\xi.$$

7.9.20. For $f(z) = z^{-1}$, verify that $f(\mathbf{A}) = \mathbf{A}^{-1}$ for every nonsingular $\mathbf{A}$.

7.9.21. If Γ is a simple closed contour enclosing all eigenvalues of a nonsingular matrix $\mathbf{A}$, what is the value of $\dfrac{1}{2\pi \mathrm{i}} \displaystyle\int_\Gamma \xi^{-1} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi$?

7.9.22. Generalized Inverses. The inverse function $f(z) = z^{-1}$ is not defined at singular matrices, but the *generalized inverse function*

$$g(z) = \begin{cases} z^{-1} & \text{if } z \neq 0, \\ 0 & \text{if } z = 0, \end{cases}$$

is defined on all square matrices. It's clear from Exercise 7.9.20 that if $\mathbf{A}$ is nonsingular, then $g(\mathbf{A}) = \mathbf{A}^{-1}$, so $g(\mathbf{A})$ is a natural way to extend the concept of inversion to include singular matrices. Explain why $g(\mathbf{A}) = \mathbf{A}^D$ is the Drazin inverse of Example 5.10.5 (p. 399) and not necessarily the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ described on p. 423.

7.9.23. Drazin Is "Natural." Suppose that $\mathbf{A}$ is a singular matrix, and let Γ be a simple closed contour that contains all eigenvalues of $\mathbf{A}$ except $\lambda_1 = 0$, which is neither in nor on Γ. Prove that

$$\frac{1}{2\pi \mathrm{i}} \int_\Gamma \xi^{-1} (\xi \mathbf{I} - \mathbf{A})^{-1} d\xi = \mathbf{A}^D$$

is the Drazin inverse for $\mathbf{A}$ as defined in Example 5.10.5 (p. 399). **Hint:** The *Cauchy–Goursat theorem* states that if a function f is analytic at all points inside and on a simple closed contour Γ, then $\int_\Gamma f(z) dz = 0$.

7.10 DIFFERENCE EQUATIONS, LIMITS, AND SUMMABILITY

A *linear difference equation* of order m with constant coefficients has the form

$$y(k+1) = \alpha_m y(k) + \alpha_{m-1} y(k-1) \cdots + \alpha_1 y(k-m+1) + \alpha_0 \tag{7.10.1}$$

in which $\alpha_0, \alpha_1, \ldots, \alpha_m$ along with initial conditions $y(0), y(1), \ldots, y(m-1)$ are known constants, and $y(m), y(m+1), y(m+2) \ldots$ are unknown. Difference equations are the discrete analogs of differential equations, and, among other ways, they arise by discretizing differential equations. For example, discretizing a second-order linear differential equation results in a system of second-order difference equations as illustrated in Example 1.4.1, p 19. The theory of linear difference equations parallels the theory for linear differential equations, and a technique similar to the one used to solve linear differential equations with constant coefficients produces the solution of (7.10.1) as

$$y(k) = \frac{\alpha_0}{1 - \alpha_1 - \cdots - \alpha_m} + \sum_{i=1}^{m} \beta_i \lambda_i^k, \qquad \text{for } k = 0, 1, \ldots \tag{7.10.2}$$

in which the λ_i 's are the roots of $\lambda^m - \alpha_m \lambda^{m-1} - \cdots - \alpha_0 = 0$, and the β_i 's are constants determined by the initial conditions $y(0), y(1), \ldots, y(m-1)$. The first term on the right-hand side of (7.10.2) is a particular solution of (7.10.1), and the summation term in (7.10.2) is the general solution of the associated homogeneous equation defined by setting $\alpha_0 = 0$.

This section focuses on systems of first-order linear difference equations with constant coefficients, and such systems can be written in matrix form as

$$\begin{aligned} &\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) && \text{(a homogeneous system)} \\ \text{or} \qquad & \\ &\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}(k) && \text{(a nonhomogeneous system),} \end{aligned} \tag{7.10.3}$$

where matrix $\mathbf{A}_{n\times n}$, the initial vector $\mathbf{x}(0)$, and vectors $\mathbf{b}(k)$, $k = 0, 1, \ldots,$ are known. The problem is to determine the unknown vectors $\mathbf{x}(k)$, $k = 1, 2, \ldots,$ along with an expression for the limiting vector $\lim_{k\to\infty} \mathbf{x}(k)$. Such systems are used to model linear discrete-time evolutionary processes, and the goal is usually to predict how (or to where) the process eventually evolves given the initial state of the process. For example, the population migration problem in Example 7.3.5 (p. 531) produces a 2×2 system of homogeneous linear difference equations (7.3.14), and the long-run (or steady-state) population distribution is obtained by finding the limiting solution. More sophisticated applications are given in Example 7.10.8 (p. 635) and Example 8.3.7 (p. 683).

Solving the equations in (7.10.3) is easy. Direct substitution verifies that

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) \quad \text{for} \quad k = 1, 2, 3, \ldots$$

and (7.10.4)

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) + \sum_{j=0}^{k-1} \mathbf{A}^{k-j-1}\mathbf{b}(j) \quad \text{for} \quad k = 1, 2, 3, \ldots$$

are respective solutions to (7.10.3). So rather than finding $\mathbf{x}(k)$ for any finite k, the real problem is to understand the nature of the limiting solution $\lim_{k\to\infty} \mathbf{x}(k)$, and this boils down to analyzing $\lim_{k\to\infty} \mathbf{A}^k$. We begin this analysis by establishing conditions under which $\mathbf{A}^k \to \mathbf{0}$.

For scalars α we know that $\alpha^k \to 0$ if and only if $|\alpha| < 1$, so it's natural to ask if there is an analogous statement for matrices. The first inclination is to replace $|\star|$ by a matrix norm $\|\star\|$, but this doesn't work for the standard norms. For example, if $\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix}$, then $\mathbf{A}^k \to \mathbf{0}$ but $\|\mathbf{A}\| = 2$ for all of the standard matrix norms. Although it's possible to construct a rather goofy-looking matrix norm $\|\star\|_g$ such that $\|\mathbf{A}\|_g < 1$ when $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$, the underlying mechanisms governing convergence to zero are better understood and analyzed by using eigenvalues and the Jordan form rather than norms. In particular, the ***spectral radius*** of $\mathbf{A}$ defined as $\rho(\mathbf{A}) = \max_{\lambda\in\sigma(\mathbf{A})} |\lambda|$ (Example 7.1.4, p. 497) plays a central role.

Convergence to Zero

For $\mathbf{A} \in \mathcal{C}^{n\times n}$, $\lim\limits_{k\to\infty} \mathbf{A}^k = \mathbf{0}$ if and only if $\rho(\mathbf{A}) < 1$. (7.10.5)

Proof. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is the Jordan form for $\mathbf{A}$, then

$$\mathbf{A}^k = \mathbf{P}\mathbf{J}^k\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star^k & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1}, \quad \text{where} \quad \mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix} \tag{7.10.6}$$

denotes a generic Jordan block in $\mathbf{J}$. Clearly, $\mathbf{A}^k \to \mathbf{0}$ if and only if $\mathbf{J}_\star^k \to \mathbf{0}$ for each Jordan block, so it suffices to prove that $\mathbf{J}_\star^k \to \mathbf{0}$ if and only if $|\lambda| < 1$. Using the function $f(z) = z^n$ in formula (7.9.2) on p. 600 along with the convention that $\binom{k}{j} = 0$ for $j > k$ produces

$$\mathbf{J}_\star^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \binom{k}{2}\lambda^{k-2} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \lambda^k & \binom{k}{1}\lambda^{k-1} & \ddots & \vdots \\ & & \ddots & \ddots & \binom{k}{2}\lambda^{k-2} \\ & & & \lambda^k & \binom{k}{1}\lambda^{k-1} \\ & & & & \lambda^k \end{pmatrix}_{m\times m}. \tag{7.10.7}$$

It's clear from the diagonal entries that if $\mathbf{J}_\star^k \to \mathbf{0}$, then $\lambda^k \to 0$, so $|\lambda| < 1$. Conversely, if $|\lambda| < 1$ then $\lim_{k\to\infty}\binom{k}{j}\lambda^{k-j} = 0$ for each fixed value of j because

$$\binom{k}{j} = \frac{k(k-1)\cdots(k-j+1)}{j!} \le \frac{k^j}{j!} \quad\Longrightarrow\quad \left|\binom{k}{j}\lambda^{k-j}\right| \le \frac{k^j}{j!}|\lambda|^{k-j} \to 0.$$

You can see that the last term on the right-hand side goes to zero as $k \to \infty$ either by applying l'Hopital's rule or by realizing that k^j goes to infinity with polynomial speed while $|\lambda|^{k-j}$ is going to zero with exponential speed. Therefore, if $|\lambda| < 1$, then $\mathbf{J}_\star^k \to \mathbf{0}$, and thus (7.10.5) is proven. ∎

Intimately related to the question of convergence to zero is the convergence of the ***Neumann series*** $\sum_{k=0}^\infty \mathbf{A}^k$. It was demonstrated in (3.8.5) on p. 126 that if $\lim_{n\to\infty}\mathbf{A}^n = \mathbf{0}$, then the Neumann series converges, and it was argued in Example 7.3.1 (p. 527) that the converse holds for diagonalizable matrices. Now we are in a position to prove that the converse is true for *all* square matrices and thereby produce the following complete statement regarding the convergence of the Neumann series.

Neumann Series

For $\mathbf{A} \in \mathcal{C}^{n\times n}$, the following statements are equivalent.

- The Neumann series $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots$ converges. (7.10.8)
- $\rho(\mathbf{A}) < 1$. (7.10.9)
- $\lim\limits_{k\to\infty} \mathbf{A}^k = \mathbf{0}$. (7.10.10)

In which case, $(\mathbf{I} - \mathbf{A})^{-1}$ exists and $\sum_{k=0}^\infty \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}$. (7.10.11)

Proof. We know from (7.10.5) that (7.10.9) and (7.10.10) are equivalent, and it was argued on p. 126 that (7.10.10) implies (7.10.8), so the theorem can be established by proving that (7.10.8) implies (7.10.9). If $\sum_{k=0}^\infty \mathbf{A}^k$ converges, it follows that $\sum_{k=0}^\infty \mathbf{J}_*^k$ must converge for each Jordan block $\mathbf{J}_*$ in the Jordan form for $\mathbf{A}$. This together with (7.10.7) implies that $\left[\sum_{k=0}^\infty \mathbf{J}_*^k\right]_{ii} = \sum_{k=0}^\infty \lambda^k$ converges for

each $\lambda \in \sigma(\mathbf{A})$, and this scalar geometric series converges if and only if $|\lambda| < 1$. Thus the convergence of $\sum_{k=0}^{\infty} \mathbf{A}^k$ implies $\rho(\mathbf{A}) < 1$. When it converges, $\sum_{k=0}^{\infty} \mathbf{A}^k = (\mathbf{I} - \mathbf{A})^{-1}$ because $(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{k-1}) = \mathbf{I} - \mathbf{A}^k \to \mathbf{I}$ as $k \to \infty$. ∎

The following examples illustrate the utility of the previous results for establishing some useful (and elegant) statements concerning spectral radius.

Example 7.10.1

Spectral Radius as a Limit. It was shown in Example 7.1.4 (p. 497) that if $\mathbf{A} \in \mathcal{C}^{n\times n}$, then $\rho(\mathbf{A}) \le \|\mathbf{A}\|$ for every matrix norm. But this was just the precursor to the following elegant relationship between spectral radius and norm.

Problem: Prove that for every matrix norm,

$$\rho(\mathbf{A}) = \lim_{k\to\infty} \left\|\mathbf{A}^k\right\|^{1/k}. \tag{7.10.12}$$

Solution: First note that $\rho(\mathbf{A})^k = \rho\left(\mathbf{A}^k\right) \le \left\|\mathbf{A}^k\right\| \implies \rho(\mathbf{A}) \le \left\|\mathbf{A}^k\right\|^{1/k}$. Next, observe that $\rho\big(\mathbf{A}/(\rho(\mathbf{A}) + \epsilon)\big) < 1$ for every $\epsilon > 0$, so, by (7.10.5),

$$\lim_{k\to\infty} \left(\frac{\mathbf{A}}{\rho(\mathbf{A}) + \epsilon}\right)^k = 0 \implies \lim_{k\to\infty} \frac{\left\|\mathbf{A}^k\right\|}{(\rho(\mathbf{A}) + \epsilon)^k} = 0.$$

Consequently, there is a positive integer K_ϵ such that $\left\|\mathbf{A}^k\right\| / (\rho(\mathbf{A}) + \epsilon)^k < 1$ for all $k \ge K_\epsilon$, so $\left\|\mathbf{A}^k\right\|^{1/k} < \rho(\mathbf{A}) + \epsilon$ for all $k \ge K_\epsilon$, and thus

$$\rho(\mathbf{A}) \le \left\|\mathbf{A}^k\right\|^{1/k} < \rho(\mathbf{A}) + \epsilon \quad \text{for } k \ge K_\epsilon.$$

Because this holds for each $\epsilon > 0$, it follows that $\lim_{k\to\infty} \left\|\mathbf{A}^k\right\|^{1/k} = \rho(\mathbf{A})$.

Example 7.10.2

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ let $|\mathbf{A}|$ denote the matrix having entries $|a_{ij}|$, and for matrices $\mathbf{B}, \mathbf{C} \in \Re^{n\times n}$ define $\mathbf{B} \le \mathbf{C}$ to mean $b_{ij} \le c_{ij}$ for each i and j.

Problem: Prove that if $|\mathbf{A}| \le \mathbf{B}$, then

$$\rho(\mathbf{A}) \le \rho(|\mathbf{A}|) \le \rho(\mathbf{B}). \tag{7.10.13}$$

Solution: The triangle inequality yields $|\mathbf{A}^k| \le |\mathbf{A}|^k$ for every positive integer k. Furthermore, $|\mathbf{A}| \le \mathbf{B}$ implies that $|\mathbf{A}|^k \le \mathbf{B}^k$. This with (7.10.12) produces

$$\begin{aligned}
\left\|\mathbf{A}^k\right\|_\infty = \left\| |\mathbf{A}^k| \right\|_\infty &\le \left\| |\mathbf{A}|^k \right\|_\infty \le \left\|\mathbf{B}^k\right\|_\infty \\
&\implies \left\|\mathbf{A}^k\right\|_\infty^{1/k} \le \left\| |\mathbf{A}|^k \right\|_\infty^{1/k} \le \left\|\mathbf{B}^k\right\|_\infty^{1/k} \\
&\implies \lim_{k\to\infty} \left\|\mathbf{A}^k\right\|_\infty^{1/k} \le \lim_{k\to\infty} \le \left\| |\mathbf{A}|^k \right\|_\infty^{1/k} \le \lim_{k\to\infty} \le \left\|\mathbf{B}^k\right\|_\infty^{1/k} \\
&\implies \rho(\mathbf{A}) \le \rho(|\mathbf{A}|) \le \rho(\mathbf{B}).
\end{aligned}$$

Example 7.10.3

Problem: Prove that if $\mathbf{0} \leq \mathbf{B}_{n\times n}$, then

$$\rho(\mathbf{B}) < r \text{ if and only if } (r\mathbf{I}-\mathbf{B})^{-1} \text{ exists and } (r\mathbf{I}-\mathbf{B})^{-1} \geq \mathbf{0}. \tag{7.10.14}$$

Solution: If $\rho(\mathbf{B}) < r$, then $\rho(\mathbf{B}/r) < 1$, so (7.10.8)–(7.10.11) imply that

$$r\mathbf{I}-\mathbf{B} = r\left(\mathbf{I}-\frac{\mathbf{B}}{r}\right) \text{ is nonsingular and } (r\mathbf{I}-\mathbf{B})^{-1} = \frac{1}{r}\sum_{k=0}^{\infty}\left(\frac{\mathbf{B}}{r}\right)^k \geq \mathbf{0}.$$

To prove the converse, it's convenient to adopt the following notation. For any $\mathbf{P} \in \Re^{m\times n}$, let $|\mathbf{P}| = \big[\,|p_{ij}|\,\big]$ denote the matrix of absolute values, and notice that the triangle inequality insures that $|\mathbf{PQ}| \leq |\mathbf{P}|\,|\mathbf{Q}|$ for all conformable $\mathbf{P}$ and $\mathbf{Q}$. Now assume that $r\mathbf{I}-\mathbf{B}$ is nonsingular and $(r\mathbf{I}-\mathbf{B})^{-1} \geq \mathbf{0}$, and prove $\rho(\mathbf{B}) < r$. Let $(\lambda, \mathbf{x})$ be any eigenpair for $\mathbf{B}$, and use $\mathbf{B} \geq \mathbf{0}$ together with $(r\mathbf{I}-\mathbf{B})^{-1} \geq \mathbf{0}$ to write

$$\begin{aligned}
\lambda\mathbf{x} = \mathbf{B}\mathbf{x} \quad &\Longrightarrow \quad |\lambda|\,|\mathbf{x}| = |\lambda\mathbf{x}| = |\mathbf{B}\mathbf{x}| \leq |\mathbf{B}|\,|\mathbf{x}| = \mathbf{B}\,|\mathbf{x}| \\
&\Longrightarrow \quad (r\mathbf{I}-\mathbf{B})|\mathbf{x}| \leq (r-|\lambda|)\,|\mathbf{x}| \\
&\Longrightarrow \quad \mathbf{0} \leq |\mathbf{x}| \leq (r-|\lambda|)\,(r\mathbf{I}-\mathbf{B})^{-1}|\mathbf{x}| \\
&\Longrightarrow \quad r-|\lambda| \geq 0.
\end{aligned} \tag{7.10.15}$$

But $|\lambda| \neq r$; otherwise (7.10.15) would imply that $|\mathbf{x}|$ (and hence $\mathbf{x}$) is zero, which is impossible. Thus $|\lambda| < r$ for all $\lambda \in \sigma(\mathbf{B})$, which means $\rho(\mathbf{B}) < r$.

Iterative algorithms are often used in lieu of direct methods to solve large sparse systems of linear equations, and some of the traditional iterative schemes fall into the following class of nonhomogeneous linear difference equations.

Linear Stationary Iterations

Let $\mathbf{Ax} = \mathbf{b}$ be a linear system that is square but otherwise arbitrary.

- A ***splitting*** of $\mathbf{A}$ is a factorization $\mathbf{A} = \mathbf{M}-\mathbf{N}$, where $\mathbf{M}^{-1}$ exists.
- Let $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ (called the ***iteration matrix***), and set $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$.
- For an initial vector $\mathbf{x}(0)_{n\times 1}$, a ***linear stationary iteration*** is

$$\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}, \qquad k = 1, 2, 3, \ldots. \tag{7.10.16}$$

- If $\rho(\mathbf{H}) < 1$, then $\mathbf{A}$ is nonsingular and

$$\lim_{k\to\infty} \mathbf{x}(k) = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad \text{for every initial vector } \mathbf{x}(0). \tag{7.10.17}$$

Proof. To prove (7.10.17), notice that if $\mathbf{A} = \mathbf{M} - \mathbf{N} = \mathbf{M}(\mathbf{I} - \mathbf{H})$ is a splitting for which $\rho(\mathbf{H}) < 1$, then (7.10.11) guarantees that $(\mathbf{I} - \mathbf{H})^{-1}$ exists, and thus $\mathbf{A}$ is nonsingular. Successive substitution applied to (7.10.16) yields

$$\mathbf{x}(k) = \mathbf{H}^k\mathbf{x}(0) + (\mathbf{I} + \mathbf{H} + \mathbf{H}^2 + \cdots + \mathbf{H}^{k-1})\mathbf{d},$$

so if $\rho(\mathbf{H}) < 1$, then (7.10.9)–(7.10.11) insures that for all $\mathbf{x}(0)$,

$$\lim_{k\to\infty} \mathbf{x}(k) = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{d} = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{M}^{-1}\mathbf{b} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x}. \quad \blacksquare \tag{7.10.18}$$

It's clear that the convergence rate of (7.10.16) is governed by the size of $\rho(\mathbf{H})$ along with the index of its associated eigenvalue (go back and look at (7.10.7)). But what really is needed is an indication of how many digits of accuracy can be expected to be gained per iteration. So as not to obscure the simple underlying idea, assume that $\mathbf{H}_{n\times n}$ is diagonalizable with

$$\sigma(\mathbf{H}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}, \quad \text{where} \quad 1 > |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_s|$$

(which is frequently the case in applications), and let $\boldsymbol{\epsilon}(k) = \mathbf{x}(k) - \mathbf{x}$ denote the error after the kth iteration. Subtracting $\mathbf{x} = \mathbf{H}\mathbf{x} + \mathbf{d}$ (a consequence of (7.10.18)) from $\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}$ produces (for large k)

$$\boldsymbol{\epsilon}(k) = \mathbf{H}\boldsymbol{\epsilon}(k-1) = \mathbf{H}^k\boldsymbol{\epsilon}(0) = (\lambda_1^k\mathbf{G}_1 + \lambda_2^k\mathbf{G}_2 + \cdots + \lambda_s^k\mathbf{G}_s)\boldsymbol{\epsilon}(0) \approx \lambda_1^k\mathbf{G}_1\boldsymbol{\epsilon}(0),$$

where the $\mathbf{G}_i$'s are the spectral projectors occurring in the spectral decomposition (pp. 517 and 520) of $\mathbf{H}^k$. Similarly, $\boldsymbol{\epsilon}(k-1) \approx \lambda_1^{k-1}\mathbf{G}_1\boldsymbol{\epsilon}(0)$, so comparing the ith components of $\boldsymbol{\epsilon}(k-1)$ and $\boldsymbol{\epsilon}(k)$ reveals that after several iterations,

$$\left|\frac{\boldsymbol{\epsilon}_i(k-1)}{\boldsymbol{\epsilon}_i(k)}\right| \approx \frac{1}{|\lambda_1|} = \frac{1}{\rho(\mathbf{H})} \quad \text{for each} \quad i = 1, 2, \ldots, n.$$

To understand the significance of this, suppose for example that

$$|\boldsymbol{\epsilon}_i(k-1)| = 10^{-q} \quad \text{and} \quad |\boldsymbol{\epsilon}_i(k)| = 10^{-p} \quad \text{with} \quad p \geq q > 0,$$

so that the error in each entry is reduced by $p - q$ digits per iteration. Since

$$p - q = \log_{10}\left|\frac{\boldsymbol{\epsilon}_i(k-1)}{\boldsymbol{\epsilon}_i(k)}\right| \approx -\log_{10}\rho(\mathbf{H}),$$

we see that $-\log_{10}\rho(\mathbf{H})$ provides us with an indication of the number of digits of accuracy that can be expected to be eventually gained on each iteration. For this reason, the number $R = -\log_{10}\rho(\mathbf{H})$ (or, alternately, $R = -\ln\rho(\mathbf{H})$) is called the ***asymptotic rate of convergence***, and this is the primary tool for comparing different linear stationary iterative algorithms.

The trick is to find splittings that guarantee rapid convergence while insuring that $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$ can be computed easily. The following three examples present the classical splittings.

Example 7.10.4

Jacobi's method[81] is produced by splitting $\mathbf{A} = \mathbf{D} - \mathbf{N}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ (we assume each $a_{ii} \neq 0$), and $-\mathbf{N}$ is the matrix containing the off-diagonal entries of $\mathbf{A}$. Clearly, both $\mathbf{H} = \mathbf{D}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{D}^{-1}\mathbf{b}$ can be formed with little effort. Notice that the ith component in the Jacobi iteration $\mathbf{x}(k) = \mathbf{D}^{-1}\mathbf{N}\mathbf{x}(k-1) + \mathbf{D}^{-1}\mathbf{b}$ is given by

$$x_i(k) = \big(b_i - \textstyle\sum_{j\neq i} a_{ij}x_j(k-1)\big)/a_{ii}. \tag{7.10.19}$$

This shows that the order in which the equations are considered is irrelevant and that the algorithm can process equations independently (or in parallel). For this reason, Jacobi's method was referred to in the 1940s as the *method of simultaneous displacements*.

Problem: Explain why Jacobi's method is guaranteed to converge for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$ when $\mathbf{A}$ is diagonally dominant as defined and discussed in Examples 4.3.3 (p. 184) and 7.1.6 (p. 499).

Solution: According to (7.10.17), it suffices to show that $\rho(\mathbf{H}) < 1$. This follows by combining $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for each i with the fact that $\rho(\mathbf{H}) \leq \|H\|_\infty$ (Example 7.1.4, p. 497) to write

$$\rho(\mathbf{H}) \leq \|H\|_\infty = \max_i \sum_j \frac{|a_{ij}|}{|a_{ii}|} = \max_i \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Example 7.10.5

The Gauss–Seidel method[82] is the result of splitting $\mathbf{A} = (\mathbf{D}-\mathbf{L})-\mathbf{U}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ ($a_{ii} \neq 0$ is assumed) and where $-\mathbf{L}$ and $-\mathbf{U}$ contain the entries occurring below and above the diagonal of $\mathbf{A}$, respectively. The iteration matrix is $\mathbf{H} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$, and $\mathbf{d} = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$. The ith entry in the Gauss–Seidel iteration $\mathbf{x}(k) = (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{x}(k-1) + (\mathbf{D}-\mathbf{L})^{-1}\mathbf{b}$ is

$$x_i(k) = \big(b_i - \textstyle\sum_{j<i} a_{ij}x_j(k) - \sum_{j>i} a_{ij}x_j(k-1)\big)/a_{ii}. \tag{7.10.20}$$

This shows that Gauss–Seidel determines $x_i(k)$ by using the newest possible information—namely, $x_1(k), x_2(k), \ldots, x_{i-1}(k)$ in the current iterate in conjunction with $x_{i+1}(k-1), x_{i+2}(k-1), \ldots, x_n(k-1)$ from the previous iterate.

---

[81] Karl Jacobi (p. 353) considered this method in 1845, but it seems to have been independently discovered by others. In addition to being called the *method of simultaneous displacements* in 1945, Jacobi's method was referred to as the *Richardson iterative method* in 1958.

[82] Ludwig Philipp von Seidel (1821–1896) studied with Dirichlet in Berlin in 1840 and with Jacobi (and others) in Königsberg. Seidel's involvement in transforming Jacobi's method into the Gauss–Seidel scheme is natural, but the reason for attaching Gauss's name is unclear. Seidel went on to earn his doctorate (1846) in Munich, where he stayed as a professor for the rest of his life. In addition to mathematics, Seidel made notable contributions in the areas of optics and astronomy, and in 1970 a lunar crater was named for Seidel.

This differs from Jacobi's method because Jacobi relies strictly on the old data in $\mathbf{x}(k-1)$. The Gauss–Seidel algorithm was known in the 1940s as the *method of successive displacements* (as opposed to the method of *simultaneous displacements*, which is Jacobi's method). Because Gauss–Seidel computes $x_i(k)$ with newer data than that used by Jacobi, it appears at first glance that Gauss–Seidel should be the superior algorithm. While this is often the case, it is not universally true—see Exercise 7.10.7.

Other Comparisons. Another major difference between Gauss–Seidel and Jacobi is that the order in which the equations are processed is irrelevant for Jacobi's method, but the value (not just the position) of the components $x_i(k)$ in the Gauss–Seidel iterate can change when the order of the equations is changed. Since this ordering feature can affect the performance of the algorithm, it was the object of much study at one time. Furthermore, when core memory is a concern, Gauss–Seidel enjoys an advantage because as soon as a new component $x_i(k)$ is computed, it can immediately replace the old value $x_i(k-1)$, whereas Jacobi requires all old values in $\mathbf{x}(k-1)$ to be retained until all new values in $\mathbf{x}(k)$ have been determined. Something that both algorithms have in common is that diagonal dominance in $\mathbf{A}$ guarantees global convergence of each method.

Problem: Explain why diagonal dominance in $\mathbf{A}$ is sufficient to guarantee convergence of the Gauss–Seidel method for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$.

Solution: Show $\rho(\mathbf{H}) < 1$. Let $(\lambda, \mathbf{z})$ be any eigenpair for $\mathbf{H}$, and suppose that the component of maximal magnitude in $\mathbf{z}$ occurs in position m. Write $(\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{z} = \lambda\mathbf{z}$ as $\lambda(\mathbf{D}-\mathbf{L})\mathbf{z} = \mathbf{U}\mathbf{z}$, and write the mth row of this latter equation as $\lambda(d-l) = u$, where

$$d = a_{mm}z_m, \quad l = -\sum_{j<m} a_{mj}z_j, \quad \text{and} \quad u = -\sum_{j>m} a_{mj}z_j.$$

Diagonal dominance $|a_{mm}| > \sum_{j\neq m}|a_{mj}|$ and $|z_j| \le |z_m|$ for all j yields

$$\begin{aligned} |u| + |l| &= \Big|\sum_{j<m} a_{mj}z_j\Big| + \Big|\sum_{j>m} a_{mj}z_j\Big| \le |z_m|\Big(\sum_{j<m}|a_{mj}| + \sum_{j>m}|a_{mj}|\Big) \\ &< |z_m||a_{mm}| = |d| \implies |u| < |d| - |l|. \end{aligned}$$

This together with $\lambda(d-l) = u$ and the backward triangle inequality (Example 5.1.1, p. 273) produces the conclusion that

$$|\lambda| = \frac{|u|}{|d-l|} \le \frac{|u|}{|d|-|l|} < 1, \quad \text{and thus} \quad \rho(\mathbf{H}) < 1.$$

Note: Diagonal dominance in $\mathbf{A}$ guarantees convergence for both Jacobi and Gauss–Seidel, but diagonal dominance is a rather severe condition that is often

not present in applications. For example the linear system in Example 7.6.2 (p. 563) that results from discretizing Laplace's equation on a square is not diagonally dominant (e.g., look at the fifth row in the 9×9 system on p. 564). But such systems are always positive definite (Example 7.6.2), and there is a classical theorem stating that *if* $\mathbf{A}$ *is positive definite, then the Gauss–Seidel iteration converges to the solution of* $\mathbf{Ax} = \mathbf{b}$ *for every initial vector* $\mathbf{x}(0)$. The same cannot be said for Jacobi's method, but there are matrices (the *M-matrices* of Example 7.10.7, p. 626) having properties resembling positive definiteness for which Jacobi's method is guaranteed to converge—see (7.10.29).

Example 7.10.6

The successive overrelaxation (SOR) method improves on Gauss–Seidel by introducing a real number $\omega \neq 0$, called a ***relaxation parameter,*** to form the splitting $\mathbf{A} = \mathbf{M} - \mathbf{N}$, where $\mathbf{M} = \omega^{-1}\mathbf{D} - \mathbf{L}$ and $\mathbf{N} = (\omega^{-1} - 1)\mathbf{D} + \mathbf{U}$. As before, $\mathbf{D}$ is the diagonal part of $\mathbf{A}$ ($a_{ii} \neq 0$ is assumed) and $-\mathbf{L}$ and $-\mathbf{U}$ contain the entries occurring below and above the diagonal of $\mathbf{A}$, respectively. Since $\mathbf{M}^{-1} = \omega(\mathbf{D} - \omega\mathbf{L})^{-1} = \omega(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}$, the SOR iteration matrix is

$$\mathbf{H}_\omega = \mathbf{M}^{-1}\mathbf{N} = (\mathbf{D} - \omega\mathbf{L})^{-1}\big[(1-\omega)\mathbf{D} + \omega\mathbf{U}\big] = (\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\big[(1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{U}\big],$$

and the kth SOR iterate emanating from (7.10.16) is

$$\mathbf{x}(k) = \mathbf{H}_\omega\mathbf{x}(k-1) + \omega(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{b}. \tag{7.10.21}$$

This is the Gauss–Seidel iteration when $\omega = 1$. Using $\omega > 1$ is called *overrelaxation*, while taking $\omega < 1$ is referred to as *underrelaxation*. Writing (7.10.21) in the form $(\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{L})\mathbf{x}(k) = \big[(1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}\mathbf{U}\big]\mathbf{x}(k-1) + \omega\mathbf{D}^{-1}\mathbf{b}$ and considering the ith component on both sides of this equality produces

$$x_i(k) = (1-\omega)x_i(k-1) + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j(k) - \sum_{j>i} a_{ij}x_j(k-1)\Big). \tag{7.10.22}$$

The matrix splitting approach is elegant and unifying, but it obscures the simple idea behind SOR. To understand the original motivation, write the Gauss–Seidel iterate in (7.10.20) as $\widetilde{x}_i(k) = \widetilde{x}_i(k-1) + c_k$, where c_k is the "correction term"

$$c_k = \frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}\widetilde{x}_j(k) - \sum_{j=i}^{n} a_{ij}\widetilde{x}_j(k-1)\Big).$$

This clearly suggests that the performance of the iteration can be affected by adjusting (or "relaxing") the correction term—i.e., by replacing c_k with ωc_k. The resulting algorithm, $\widetilde{x}_i(k) = \widetilde{x}_i(k-1) + \omega c_k$, is in fact (7.10.22), which produces (7.10.21). Moreover, it was observed early on that Gauss–Seidel applied to finite difference approximations for elliptic partial differential equations, such

as the one in Example 7.6.2 (p. 563), often produces successive corrections c_k that have the same sign, so it was reasoned that convergence might be accelerated for these applications by increasing the magnitude of the correction factor at each step (i.e., by setting $\omega > 1$). Thus the technique became known as "successive overrelaxation" rather than simply "successive relaxation." It's not hard to see that $\rho(\mathbf{H}_\omega) < 1$ only if $0 < \omega < 2$ (Exercise 7.10.9), and it can be proven that positive definiteness of $\mathbf{A}$ is sufficient to guarantee $\rho(\mathbf{H}_\omega) < 1$ whenever $0 < \omega < 2$. But determining ω to minimize $\rho(\mathbf{H}_\omega)$ is generally a difficult task.

Nevertheless, there is one famous special case[83] for which the optimal value of ω can be explicitly given. If $\det(\alpha\mathbf{D} - \mathbf{L} - \mathbf{U}) = \det\left(\alpha\mathbf{D} - \beta\mathbf{L} - \beta^{-1}\mathbf{U}\right)$ for all real α and $\beta \neq 0$, and if the iteration matrix $\mathbf{H}_J$ for Jacobi's method has real eigenvalues with $\rho(\mathbf{H}_J) < 1$, then the eigenvalues λ_J for $\mathbf{H}_J$ are related to the eigenvalues λ_ω of $\mathbf{H}_\omega$ by

$$(\lambda_\omega + \omega - 1)^2 = \omega^2 \lambda_J^2 \lambda_\omega. \tag{7.10.23}$$

From this it can be proven that the optimum value of ω for SOR is

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{H}_J)}} \quad \text{and} \quad \rho\left(\mathbf{H}_{\omega_{\text{opt}}}\right) = \omega_{\text{opt}} - 1. \tag{7.10.24}$$

Furthermore, setting $\omega = 1$ in (7.10.23) yields $\rho(\mathbf{H}_{GS}) = \rho^2(\mathbf{H}_J)$, where $\mathbf{H}_{GS}$ is the Gauss–Seidel iteration matrix. For example, the discrete Laplacian $\mathbf{L}_{n^2\times n^2}$ in Example 7.6.2 (p. 563) satisfies the special case conditions, and the spectral radii of the iteration matrices associated with $\mathbf{L}$ are

$$\begin{aligned}
\text{Jacobi:} \quad \rho(\mathbf{H}_J) &= \cos\pi h && \approx 1 - (\pi^2 h^2/2) \quad \text{(see Exercise 7.10.10)},\\
\text{Gauss–Seidel:} \quad \rho(\mathbf{H}_{GS}) &= \cos^2\pi h && \approx 1 - \pi^2 h^2,\\
\text{SOR:} \quad \rho\left(\mathbf{H}_{\omega_{\text{opt}}}\right) &= \frac{1 - \sin\pi h}{1 + \sin\pi h} && \approx 1 - 2\pi h,
\end{aligned}$$

where we have set $h = 1/(n+1)$. Examining asymptotic rates of convergence reveals that Gauss–Seidel is twice as fast as Jacobi on the discrete Laplacian because $R_{GS} = -\log_{10}\cos^2\pi h = -2\log_{10}\cos\pi h = 2R_J$. However, optimal SOR is much better because $1 - 2\pi h$ is significantly smaller than $1 - \pi^2 h^2$ for even moderately small h. The point is driven home by looking at the asymptotic rates of convergence for $h = .02$ $(n = 49)$ as shown below:

$$\begin{aligned}
\text{Jacobi:} \quad & R_J \approx .000858,\\
\text{Gauss–Seidel:} \quad & R_{GS} = 2R_J \approx .001716,\\
\text{SOR:} \quad & R_{\text{opt}} \approx .054611 \approx 32R_{GS} = 64R_J.
\end{aligned}$$

[83] This special case was developed by the contemporary numerical analyst David M. Young, Jr., who produced much of the SOR theory in his 1950 Ph.D. dissertation that was directed by Garrett Birkhoff at Harvard University. The development of SOR is considered to be one of the major computational achievements of the first half of the twentieth century, and it motivated at least two decades of intense effort in matrix computations.

In other words, after things settle down, a single SOR step on $\mathbf{L}$ (for $h = .02$) is equivalent to about 32 Gauss–Seidel steps and 64 Jacobi steps!

Note: In spite of the preceding remarks, SOR has limitations. Special cases for which the optimum ω can be explicitly determined are rare, so adaptive computational procedures are generally necessary to approximate a good ω, and the results are often not satisfying. While SOR was a big step forward over the algorithms of the nineteenth century, the second half of the twentieth century saw the development of more robust methods—such as the preconditioned conjugate gradient method (p. 657) and GMRES (p. 655)—that have relegated SOR to a secondary role.

Example 7.10.7

M-matrices[84] are real nonsingular matrices $\mathbf{A}_{n\times n}$ such that $a_{ij} \leq 0$ for all $i \neq j$ and $\mathbf{A}^{-1} \geq \mathbf{0}$ (each entry of $\mathbf{A}^{-1}$ is nonnegative). They arise naturally in a broad variety of applications ranging from economics (Example 8.3.6, p. 681) to hard-core engineering problems, and, as shown in (7.10.29), they are particularly relevant in formulating and analyzing iterative methods. Some important properties of M-matrices are developed below.

- $\mathbf{A}$ is an M-matrix if and only if there exists a matrix $\mathbf{B} \geq \mathbf{0}$ and a real number $r > \rho(\mathbf{B})$ such that $\mathbf{A} = r\mathbf{I} - \mathbf{B}$. (7.10.25)

- If $\mathbf{A}$ is an M-matrix, then $\operatorname{Re}(\lambda) > 0$ for all $\lambda \in \sigma(\mathbf{A})$. Conversely, all matrices with nonpositive off-diagonal entries whose spectrums are in the right-hand halfplane are M-matrices. (7.10.26)

- Principal submatrices of M-matrices are also M-matrices. (7.10.27)

- If $\mathbf{A}$ is an M-matrix, then all principal minors in $\mathbf{A}$ are positive. Conversely, all matrices with nonpositive off-diagonal entries whose principal minors are positive are M-matrices. (7.10.28)

- If $\mathbf{A} = \mathbf{M} - \mathbf{N}$ is a splitting of an M-matrix for which $\mathbf{M}^{-1} \geq \mathbf{0}$, then the linear stationary iteration (7.10.16) is convergent for all initial vectors $\mathbf{x}(0)$ and for all right-hand sides $\mathbf{b}$. In particular, Jacobi's method in Example 7.10.4 (p. 622) converges for all M-matrices. (7.10.29)

Proof of (7.10.25). Suppose that $\mathbf{A}$ is an M-matrix, and let $r = \max_i |a_{ii}|$ so that $\mathbf{B} = r\mathbf{I} - \mathbf{A} \geq \mathbf{0}$. Since $\mathbf{A}^{-1} = (r\mathbf{I} - \mathbf{B})^{-1} \geq \mathbf{0}$, it follows from (7.10.14) in Example 7.10.3 (p. 620) that $r > \rho(\mathbf{B})$. Conversely, if $\mathbf{A}$ is any matrix of

[84] This terminology was introduced in 1937 by the twentieth-century mathematician Alexander Markowic Ostrowski, who made several contributions to the analysis of classical iterative methods. The "M" is short for "Minkowski" (p. 278).

the form $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, then (7.10.14) guarantees that $\mathbf{A}^{-1}$ exists and $\mathbf{A}^{-1} \geq \mathbf{0}$, and it's clear that $a_{ij} \leq 0$ for each $i \neq j$, so $\mathbf{A}$ must be an M-matrix. ∎

Proof of (7.10.26). If $\mathbf{A}$ is an M-matrix, then, by (7.10.25), $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $r > \rho(\mathbf{B})$. This means that if $\lambda_{\mathbf{A}} \in \sigma(\mathbf{A})$, then $\lambda_{\mathbf{A}} = r - \lambda_{\mathbf{B}}$ for some $\lambda_{\mathbf{B}} \in \sigma(\mathbf{B})$. If $\lambda_{\mathbf{B}} = \alpha + \mathrm{i}\beta$, then $r > \rho(\mathbf{B}) \geq |\lambda_{\mathbf{B}}| = \sqrt{\alpha^2 + \beta^2} \geq |\alpha| \geq \alpha$ implies that $\mathrm{Re}(\lambda_{\mathbf{A}}) = r - \alpha \geq 0$. Now suppose that $\mathbf{A}$ is any matrix such that $a_{ij} \leq 0$ for all $i \neq j$ and $\mathrm{Re}(\lambda_{\mathbf{A}}) > 0$ for all $\lambda_{\mathbf{A}} \in \sigma(\mathbf{A})$. This means that there is a real number γ such that the circle centered at γ and having radius equal to γ contains $\sigma(\mathbf{A})$—see Figure 7.10.1. Let r be any real number such that $r > \max\{2\gamma, \max_i |a_{ii}|\}$, and set $\mathbf{B} = r\mathbf{I} - \mathbf{A}$. It's apparent that $\mathbf{B} \geq \mathbf{0}$, and, as can be seen from Figure 7.10.1, the distance $|r - \lambda_{\mathbf{A}}|$ between r and every point in $\sigma(\mathbf{A})$ is less than r.

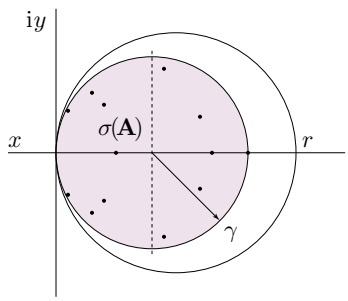


FIGURE 7.10.1

All eigenvalues of $\mathbf{B}$ look like $\lambda_{\mathbf{B}} = r - \lambda_{\mathbf{A}}$, and $|\lambda_{\mathbf{B}}| = |r - \lambda_{\mathbf{A}}| < r$, so $\rho(\mathbf{B}) < r$. Since $\mathbf{A} = r\mathbf{I} - \mathbf{B}$ is nonsingular (because $0 \notin \sigma(\mathbf{A})$) with $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, it follows from (7.10.14) in Example 7.10.3 (p. 620) that $\mathbf{A}^{-1} \geq \mathbf{0}$, and thus $\mathbf{A}$ is an M-matrix. ∎

Proof of (7.10.27). If $\widetilde{\mathbf{A}}_{k\times k}$ is the principal submatrix lying on the intersection of rows and columns $i_1, \ldots, i_k$ in an M-matrix $\mathbf{A} = r\mathbf{I} - \mathbf{B}$, where $\mathbf{B} \geq \mathbf{0}$ and $r > \rho(\mathbf{B})$, then $\widetilde{\mathbf{A}} = r\mathbf{I} - \widetilde{\mathbf{B}}$, where $\widetilde{\mathbf{B}} \geq \mathbf{0}$ is the corresponding principal submatrix of $\mathbf{B}$. Let $\mathbf{P}$ be a permutation matrix such that

$$\mathbf{P}^T\mathbf{B}\mathbf{P} = \begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}, \text{ or } \mathbf{B} = \mathbf{P}\begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}\mathbf{P}^T, \text{ and let } \mathbf{C} = \mathbf{P}\begin{pmatrix} \widetilde{\mathbf{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^T.$$

Clearly, $\mathbf{0} \leq \mathbf{C} \leq \mathbf{B}$, so, by (7.10.13) on p. 619, $\rho(\widetilde{\mathbf{B}}) = \rho(\mathbf{C}) \leq \rho(\mathbf{B}) < r$. Consequently, (7.10.25) insures that $\widetilde{\mathbf{A}}$ is an M-matrix. ∎

Proof of (7.10.28). If $\mathbf{A}$ is an M-matrix, then $\det(\mathbf{A}) > 0$ because the eigenvalues of a real matrix appear in complex conjugate pairs, so (7.10.26) and (7.1.8),

p. 494, guarantee that $\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i > 0$. It follows that each principal minor is positive because each submatrix of an M-matrix is again an M-matrix. Now prove that if $\mathbf{A}_{n\times n}$ is a matrix such that $a_{ij} \le 0$ for $i \ne j$ and each principal minor is positive, then $\mathbf{A}$ must be an M-matrix. Proceed by induction on n. For $n = 1$, the assumption of positive principal minors implies that $\mathbf{A} = [\rho]$ with $\rho > 0$, so $\mathbf{A}^{-1} = 1/\rho > 0$. Suppose the result is true for $n = k$, and consider the LU factorization

$$\mathbf{A}_{(k+1)\times(k+1)} = \begin{pmatrix} \widetilde{\mathbf{A}}_{k\times k} & \mathbf{c} \\ \mathbf{d}^T & \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{d}^T\widetilde{\mathbf{A}}^{-1} & 1 \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{c} \\ \mathbf{0} & \alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c} \end{pmatrix} = \mathbf{LU}.$$

We know that $\mathbf{A}$ is nonsingular ($\det(\mathbf{A})$ is a principal minor) and $\alpha > 0$ (it's a 1×1 principal minor), and the induction hypothesis insures that $\widetilde{\mathbf{A}}^{-1} \ge \mathbf{0}$. Combining these facts with $\mathbf{c} \le \mathbf{0}$ and $\mathbf{d}^T \le \mathbf{0}$ produces

$$\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1} = \begin{pmatrix} \widetilde{\mathbf{A}}^{-1} & \dfrac{-\widetilde{\mathbf{A}}^{-1}\mathbf{c}}{\alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c}} \\ \mathbf{0} & \dfrac{1}{\alpha - \mathbf{d}^T\widetilde{\mathbf{A}}^{-1}\mathbf{c}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{d}^T\widetilde{\mathbf{A}}^{-1} & 1 \end{pmatrix} \ge \mathbf{0},$$

and thus the induction argument is completed. ∎

Proof of (7.10.29). If $\mathbf{A} = \mathbf{M} - \mathbf{N}$ is an M-matrix, and if $\mathbf{M}^{-1} \ge \mathbf{0}$ and $\mathbf{N} \ge \mathbf{0}$, then the iteration matrix $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ is clearly nonnegative. Furthermore,

$$(\mathbf{I} - \mathbf{H})^{-1} - \mathbf{I} = (\mathbf{I} - \mathbf{H})^{-1}\mathbf{H} = \mathbf{A}^{-1}\mathbf{N} \ge \mathbf{0} \implies (\mathbf{I} - \mathbf{H})^{-1} \ge \mathbf{I} \ge \mathbf{0},$$

so (7.10.14) in Example 7.10.3 (p. 620) insures that $\rho(\mathbf{H}) < 1$. Convergence of Jacobi's method is a special case because the Jacobi splitting is $\mathbf{A} = \mathbf{D} - \mathbf{N}$, where $\mathbf{D} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$, and (7.10.28) implies that each $a_{ii} > 0$. ∎

Note: Comparing properties of M-matrices with those of positive definite matrices reveals many parallels, and, in a rough sense, an M-matrix often plays the role of "a poor man's positive definite matrix." Only a small sample of M-matrix theory has been presented here, but there is in fact enough to fill a monograph on the subject. For example, there are at least 50 known equivalent conditions that can be imposed on a real matrix with nonpositive off-diagonal entries (often called a *Z-matrix*) to guarantee that it is an M-matrix—see Exercise 7.10.12 for a sample of such conditions in addition to those listed above.

We now focus on broader issues concerning when $\lim_{k\to\infty}\mathbf{A}^k$ exists but may be nonzero. Start from the fact that $\lim_{k\to\infty}\mathbf{A}^k$ exists if and only if $\lim_{k\to\infty}\mathbf{J}_\star^k$ exists for each Jordan block in (7.10.6). It's clear from (7.10.7) that $\lim_{k\to\infty}\mathbf{J}_\star^k$ cannot exist when $|\lambda|>1$, and we already know the story for $|\lambda|<1$, so we only have to examine the case when $|\lambda|=1$. If $|\lambda|=1$ with $\lambda\neq 1$ (i.e., $\lambda=\mathrm{e}^{\mathrm{i}\theta}$ with $0<\theta<2\pi$), then the diagonal terms λ^k oscillate indefinitely, and this prevents $\mathbf{J}_\star^k$ (and $\mathbf{A}^k$) from having a limit. When $\lambda=1$,

$$\mathbf{J}_\star^k=\begin{pmatrix}1 & \binom{k}{1} & \cdots & \binom{k}{m-1}\\ & \ddots & \ddots & \vdots\\ & & \ddots & \binom{k}{1}\\ & & & 1\end{pmatrix}_{m\times m} \tag{7.10.30}$$

has a limiting value if and only if $m=1$, which is equivalent to saying that $\lambda=1$ is a semisimple eigenvalue. But $\lambda=1$ may be repeated p times so that there are p Jordan blocks of the form $\mathbf{J}_\star=[1]_{1\times 1}$. Consequently, $\lim_{k\to\infty}\mathbf{A}^k$ exists if and only if the Jordan form for $\mathbf{A}$ has the structure

$$\mathbf{J}=\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}\end{pmatrix},\ \text{where } p=\mathit{alg\ mult}(1)\ \text{ and }\ \rho(\mathbf{K})<1. \tag{7.10.31}$$

Now that we know when $\lim_{k\to\infty}\mathbf{A}^k$ exists, let's describe what $\lim_{k\to\infty}\mathbf{A}^k$ looks like. We already know the answer when $p=0$—it's $\mathbf{0}$ (because $\rho(\mathbf{A})<1$). But when p is nonzero, $\lim_{k\to\infty}\mathbf{A}^k\neq\mathbf{0}$, and it can be evaluated in a couple of different ways. One way is to partition $\mathbf{P}=\begin{pmatrix}\mathbf{P}_1\,|\,\mathbf{P}_2\end{pmatrix}$ and $\mathbf{P}^{-1}=\begin{pmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{pmatrix}$, and use (7.10.5) and (7.10.31) to write

$$\begin{aligned}\lim_{k\to\infty}\mathbf{A}^k_{n\times n}&=\lim_{k\to\infty}\mathbf{P}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{K}^k\end{pmatrix}\mathbf{P}^{-1}=\mathbf{P}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{P}^{-1}\\ &=\begin{pmatrix}\mathbf{P}_1\,|\,\mathbf{P}_2\end{pmatrix}\begin{pmatrix}\mathbf{I}_{p\times p} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Q}_1\\ \mathbf{Q}_2\end{pmatrix}=\mathbf{P}_1\mathbf{Q}_1=\mathbf{G}.\end{aligned} \tag{7.10.32}$$

Another way is to use $f(z)=z^k$ in the spectral resolution theorem on p. 603. If $\sigma(\mathbf{A})=\{\lambda_1,\lambda_2,\ldots,\lambda_s\}$ with $1=\lambda_1>|\lambda_2|\geq\cdots\geq|\lambda_s|$, and if $\mathit{index}(\lambda_i)=k_i$, where $k_1=1$, then $\lim_{k\to\infty}\binom{k}{j}\lambda_i^{k-j}=0$ for $i\geq 2$ (see p. 618), and

$$\begin{aligned}\mathbf{A}^k&=\sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\binom{k}{j}\lambda_i^{k-j}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i\\ &=\mathbf{G}_1+\sum_{i=2}^{s}\sum_{j=0}^{k_i-1}\binom{k}{j}\lambda_i^{k-j}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i\to\mathbf{G}_1\quad\text{as}\quad k\to\infty.\end{aligned}$$

In other words, $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{G}_1 = \mathbf{G}$ is the spectral projector associated with $\lambda_1 = 1$. Since $index(\lambda_1) = 1$, we know from the discussion on p. 603 that $R(\mathbf{G}) = N(\mathbf{I} - \mathbf{A})$ and $N(\mathbf{G}) = R(\mathbf{I} - \mathbf{A})$. Notice that if $\rho(\mathbf{A}) < 1$, then $\mathbf{I} - \mathbf{A}$ is nonsingular, and $N(\mathbf{I} - \mathbf{A}) = \{\mathbf{0}\}$. So regardless of whether the limit is zero or nonzero, $\lim_{k\to\infty} \mathbf{A}^k$ is always the projector onto $N(\mathbf{I} - \mathbf{A})$ along $R(\mathbf{I} - \mathbf{A})$. Below is a summary of the above observations.

Limits of Powers

For $\mathbf{A} \in \mathcal{C}^{n\times n}$, $\lim_{k\to\infty} \mathbf{A}^k$ exists if and only if

$$\rho(\mathbf{A}) < 1$$

or else (7.10.33)

$$\rho(\mathbf{A}) = 1, \quad \text{where } \lambda = 1 \text{ is the only eigenvalue on the unit circle, and } \lambda = 1 \text{ is semisimple.}$$

When it exists,

$$\lim_{k\to\infty} \mathbf{A}^k = \text{the projector onto } N(\mathbf{I} - \mathbf{A}) \text{ along } R(\mathbf{I} - \mathbf{A}). \tag{7.10.34}$$

With each scalar sequence $\{\alpha_1, \alpha_2, \alpha_3, \ldots\}$ there is an associated sequence of averages $\{\mu_1, \mu_2, \mu_3, \ldots\}$ in which

$$\mu_1 = \alpha_1, \quad \mu_2 = \frac{\alpha_1 + \alpha_2}{2}, \quad \ldots, \quad \mu_n = \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_n}{n}.$$

This sequence of averages is called the associated ***Cesàro sequence***,[85] and when $\lim_{n\to\infty} \mu_n = \alpha$, we say that $\{\alpha_n\}$ is ***Cesàro summable*** (or merely *summable*) to α. It can be proven (Exercise 7.10.11) that if $\{\alpha_n\}$ converges to α, then $\{\mu_n\}$ converges to α, but not conversely. In other words, convergence implies summability, but summability doesn't insure convergence. To see that a sequence can be summable without being convergent, notice that the oscillatory sequence $\{0, 1, 0, 1, \ldots\}$ doesn't converge, but it is Cesàro summable to $1/2$, which is the mean value of $\{0, 1\}$. This is typical because averaging has a smoothing effect so that oscillations that prohibit convergence of the original sequence tend to be smoothed away or averaged out in the Cesàro sequence.

[85] Ernesto Cesàro (1859–1906) was an Italian mathematician who worked mainly in differential geometry but also contributed to number theory, divergent series, and mathematical physics. After studying in Naples, Liège, and Paris, Cesàro received his doctorate from the University of Rome in 1887, and he went on to occupy the chair of mathematics at Palermo. Cesàro's most important contribution is considered to be his 1890 book *Lezione di geometria intrinseca*, but, in large part, his name has been perpetuated because of its attachment to the concept of Cesàro summability.

Similar statements hold for general sequences of vectors and matrices (Exercise 7.10.11), but Cesàro summability is particularly interesting when it is applied to the sequence $\mathcal{P} = \{\mathbf{A}^k\}_{k=0}^{\infty}$ of powers of a square matrix $\mathbf{A}$. We know from (7.10.33) and (7.10.34) under what conditions sequence $\mathcal{P}$ converges as well as the nature of the limit, so let's now suppose that $\mathcal{P}$ doesn't converge, and decide when $\mathcal{P}$ is summable, and what $\mathcal{P}$ is summable to.

From now on, we will say that $\mathbf{A}_{n\times n}$ is a ***convergent matrix*** when $\lim_{k\to\infty}\mathbf{A}^k$ exists, and we will say that $\mathbf{A}$ is a ***summable matrix*** when $\lim_{k\to\infty}(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{k-1})/k$ exists. As in the scalar case, if $\mathbf{A}$ is convergent to $\mathbf{G}$, then $\mathbf{A}$ is summable to $\mathbf{G}$, but not conversely (Exercise 7.10.11). To analyze the summability of $\mathbf{A}$ in the absence of convergence, begin with the observation that $\mathbf{A}$ is summable if and only if the Jordan form $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ for $\mathbf{A}$ is summable, which in turn is equivalent to saying that each Jordan block $\mathbf{J}_\star$ in $\mathbf{J}$ is summable. Consequently, $\mathbf{A}$ cannot be summable whenever $\rho(\mathbf{A}) > 1$ because if $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ is a Jordan block in which $|\lambda| > 1$, then each diagonal entry of $\left(\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}\right)/k$ is

$$\delta(\lambda,k) = \frac{1+\lambda+\cdots+\lambda^{k-1}}{k} = \frac{1}{k}\left(\frac{1-\lambda^k}{1-\lambda}\right) = \frac{1}{1-\lambda}\left(\frac{1}{k}-\frac{\lambda^k}{k}\right), \qquad (7.10.35)$$

and this becomes unbounded as $k \to \infty$. In other words, it's necessary that $\rho(\mathbf{A}) \le 1$ for $\mathbf{A}$ to be summable. Since we already know that $\mathbf{A}$ is convergent (and hence summable) to $\mathbf{0}$ when $\rho(\mathbf{A}) < 1$, we need only consider the case when $\mathbf{A}$ has eigenvalues on the unit circle.

If $\lambda \in \sigma(\mathbf{A})$ such that $|\lambda| = 1$, $\lambda \neq 1$, and if $\mathit{index}(\lambda) > 1$, then there is an associated Jordan block $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ that is larger than 1×1. Each entry on the first superdiagonal of $\left(\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}\right)/k$ is the derivative $\partial\delta/\partial\lambda$ of the expression in (7.10.35), and it's not hard to see that $\partial\delta/\partial\lambda$ oscillates indefinitely as $k \to \infty$. In other words, $\mathbf{A}$ cannot be summable if there are eigenvalues $\lambda \neq 1$ on the unit circle such that $\mathit{index}(\lambda) > 1$.

Similarly, if $\lambda = 1$ is an eigenvalue of index greater than one, then $\mathbf{A}$ can't be summable because each entry on the first superdiagonal of

$$\frac{\mathbf{I}+\mathbf{J}_\star+\cdots+\mathbf{J}_\star^{k-1}}{k} \quad \text{is} \quad \frac{1+2+\cdots+(k-1)}{k} = \frac{k(k-1)}{2k} = \frac{k-1}{2} \to \infty.$$

Therefore, if $\mathbf{A}$ is summable and has eigenvalues λ such that $|\lambda| = 1$, then it's necessary that $\mathit{index}(\lambda) = 1$. The condition also is sufficient—i.e., if $\rho(\mathbf{A}) = 1$ and each eigenvalue on the unit circle is semisimple, then $\mathbf{A}$ is summable. This follows because each Jordan block associated with an eigenvalue μ such that $|\mu| < 1$ is convergent (and hence summable) to $\mathbf{0}$ by (7.10.5), and for semisimple

eigenvalues λ such that $|\lambda| = 1$, the associated Jordan blocks are 1×1 and hence summable because (7.10.35) implies

$$\frac{1+\lambda+\cdots+\lambda^{k-1}}{k} = \begin{cases} \dfrac{1}{1-\lambda}\left(\dfrac{1}{k} - \dfrac{\lambda^k}{k}\right) \to 0 & \text{for } |\lambda| = 1,\ \lambda \neq 1, \\ 1 & \text{for } \lambda = 1. \end{cases}$$

In addition to providing a necessary and sufficient condition for $\mathbf{A}$ to be Cesàro summable, the preceding analysis also reveals the nature of the Cesàro limit because if $\mathbf{A}$ is summable, then each Jordan block $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{pmatrix}$ in the Jordan form for $\mathbf{A}$ is summable, in which case we have established that

$$\lim_{k\to\infty} \frac{\mathbf{I} + \mathbf{J}_\star + \cdots + \mathbf{J}_\star^{k-1}}{k} = \begin{cases} \big[1\big]_{1\times 1} & \text{if } \lambda = 1 \text{ and } \textit{index}(\lambda) = 1, \\ \big[0\big]_{1\times 1} & \text{if } |\lambda| = 1,\ \lambda \neq 1, \text{ and } \textit{index}(\lambda) = 1, \\ \mathbf{0} & \text{if } |\lambda| < 1. \end{cases}$$

Consequently, if $\mathbf{A}$ is summable, then the Jordan form for $\mathbf{A}$ must look like

$$\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \quad \text{where} \quad p = \textit{alg mult}_{\mathbf{A}}\,(\lambda = 1)\,,$$

and the eigenvalues of $\mathbf{C}$ are such that $|\lambda| < 1$ or else $|\lambda| = 1$, $\lambda \neq 1$, $\textit{index}(\lambda) = 1$. So $\mathbf{C}$ is summable to $\mathbf{0}$, $\mathbf{J}$ is summable to $\begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, and

$$\frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k} = \mathbf{P}\left(\frac{\mathbf{I} + \mathbf{J} + \cdots + \mathbf{J}^{k-1}}{k}\right)\mathbf{P}^{-1} \to \mathbf{P}\begin{pmatrix} \mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^{-1} = \mathbf{G}.$$

Comparing this expression with that in (7.10.32) reveals that *the Cesàro limit is exactly the same as the ordinary limit, had it existed.* In other words, if $\mathbf{A}$ is summable, then regardless of whether or not $\mathbf{A}$ is convergent, $\mathbf{A}$ is summable to the projector onto $N\,(\mathbf{I} - \mathbf{A})$ along $R\,(\mathbf{I} - \mathbf{A})$. Below is a formal summary of our observations concerning Cesàro summability.

Cesàro Summability

- $\mathbf{A} \in \mathcal{C}^{n\times n}$ is Cesàro summable if and only if $\rho(\mathbf{A}) < 1$ or else $\rho(\mathbf{A}) = 1$ with each eigenvalue on the unit circle being semisimple.
- When it exists, the Cesàro limit

$$\lim_{k\to\infty} \frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k} = \mathbf{G} \tag{7.10.36}$$

  is the projector onto $N(\mathbf{I} - \mathbf{A})$ along $R(\mathbf{I} - \mathbf{A})$.
- $\mathbf{G} \neq \mathbf{0}$ if and only if $1 \in \sigma(\mathbf{A})$, in which case $\mathbf{G}$ is the spectral projector associated with $\lambda = 1$.
- If $\mathbf{A}$ is convergent to $\mathbf{G}$, then $\mathbf{A}$ is summable to $\mathbf{G}$, but not conversely.

Since the projector $\mathbf{G}$ onto $N(\mathbf{I} - \mathbf{A})$ along $R(\mathbf{I} - \mathbf{A})$ plays a prominent role, let's consider how $\mathbf{G}$ might be computed. Of course, we could just iterate on $\mathbf{A}^k$ or $(\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1})/k$, but this is inefficient and, depending on the proximity of the eigenvalues relative to the unit circle, convergence can be slow—averaging in particular can be extremely slow. The Jordan form is the basis for the theoretical development, but using it to compute $\mathbf{G}$ would be silly (see p. 592). The formula for a projector given in (5.9.12) on p. 386 is a possibility, but using a full-rank factorization of $\mathbf{I} - \mathbf{A}$ is an attractive alternative.

A ***full-rank factorization*** of a matrix $\mathbf{M}_{m\times n}$ of rank r is a factorization

$$\mathbf{M} = \mathbf{B}_{m\times r}\mathbf{C}_{r\times n}, \quad \text{where} \quad rank\,(\mathbf{B}) = rank\,(\mathbf{C}) = r = rank\,(\mathbf{M}). \tag{7.10.37}$$

All of the standard reduction techniques produce full-rank factorizations. For example, Gaussian elimination can be used because if $\mathbf{B}$ is the matrix of basic columns of $\mathbf{M}$, and if $\mathbf{C}$ is the matrix containing the nonzero rows in the reduced row echelon form $\mathbf{E}_{\mathbf{M}}$, then $\mathbf{M} = \mathbf{BC}$ is a full-rank factorization (Exercise 3.9.8, p. 140). If orthogonal reduction (p. 341) is used to produce a unitary matrix $\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 \\ \hline \mathbf{P}_2 \end{pmatrix}$ and an upper-trapezoidal matrix $\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 \\ \hline \mathbf{0} \end{pmatrix}$ such that $\mathbf{PA} = \mathbf{T}$, where $\mathbf{P}_1$ is $r \times m$ and $\mathbf{T}_1$ contains the nonzero rows, then $\mathbf{M} = \mathbf{P}_1^*\mathbf{T}_1$ is a full-rank factorization. If

$$\mathbf{M} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^* = (\mathbf{U}_1 \,|\, \mathbf{U}_2)\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{V}_1^* \\ \hline \mathbf{V}_2^* \end{pmatrix} = \mathbf{U}_1\mathbf{D}\mathbf{V}_1^* \tag{7.10.38}$$

is the singular value decomposition (5.12.2) on p. 412 (a URV factorization (p. 407) could also be used), then $\mathbf{M} = \mathbf{U}_1(\mathbf{D}\mathbf{V}_1^*) = (\mathbf{U}_1\mathbf{D})\mathbf{V}_1^*$ are full-rank factorizations. Projectors, in general, and limiting projectors, in particular, are nicely described in terms of full-rank factorizations.

Projectors

If $\mathbf{M}_{n\times n} = \mathbf{B}_{n\times r}\mathbf{C}_{r\times n}$ is any full-rank factorization as described in (7.10.37), and if $R(\mathbf{M})$ and $N(\mathbf{M})$ are complementary subspaces of $\mathcal{C}^n$, then the projector onto $R(\mathbf{M})$ along $N(\mathbf{M})$ is given by

$$\mathbf{P} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C} \tag{7.10.39}$$

or

$$\mathbf{P} = \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^* \quad \text{when (7.10.38) is used.} \tag{7.10.40}$$

If $\mathbf{A}$ is convergent or summable to $\mathbf{G}$ as described in (7.10.34) and (7.10.36), and if $\mathbf{I} - \mathbf{A} = \mathbf{B}\mathbf{C}$ is a full-rank factorization, then

$$\mathbf{G} = \mathbf{I} - \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C} \tag{7.10.41}$$

or

$$\mathbf{G} = \mathbf{I} - \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^* \quad \text{when (7.10.38) is used.} \tag{7.10.42}$$

Note: Formulas (7.10.39) and (7.10.40) are extensions of (5.13.3) on p. 430.

Proof. It's always true (Exercise 4.5.12, p. 220) that

$$\begin{aligned} R(\mathbf{X}_{m\times n}\mathbf{Y}_{n\times p}) &= R(\mathbf{X}) \quad \text{when} \quad rank(\mathbf{Y}) = n, \\ N(\mathbf{X}_{m\times n}\mathbf{Y}_{n\times p}) &= N(\mathbf{Y}) \quad \text{when} \quad rank(\mathbf{X}) = n. \end{aligned} \tag{7.10.43}$$

If $\mathbf{M}_{n\times n} = \mathbf{B}_{n\times r}\mathbf{C}_{r\times n}$ is a full-rank factorization, and if $R(\mathbf{M})$ and $N(\mathbf{M})$ are complementary subspaces of $\mathcal{C}^N$, then $rank(\mathbf{M}) = rank(\mathbf{M}^2)$ (Exercise 5.10.12, p. 402), so combining this with the first part of (7.10.43) produces

$$r = rank(\mathbf{B}\mathbf{C}) = rank(\mathbf{B}\mathbf{C}\mathbf{B}\mathbf{C}) = rank(\mathbf{C}\mathbf{B})_{r\times r} \implies (\mathbf{C}\mathbf{B})^{-1} \text{ exists.}$$

$\mathbf{P} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{C}$ is a projector because $\mathbf{P}^2 = \mathbf{P}$ (recall (5.9.8), p. 386), and (7.10.43) insures that $R(\mathbf{P}) = R(\mathbf{B}) = R(\mathbf{M})$ and $N(\mathbf{P}) = N(\mathbf{C}) = N(\mathbf{M})$. Thus (7.10.39) is proved. If (7.10.38) is used to produce a full-rank factorization $\mathbf{M} = \mathbf{U}_1(\mathbf{D}\mathbf{V}_1^*)$, then, because $\mathbf{D}$ is nonsingular,

$$\mathbf{P} = (\mathbf{U}_1\mathbf{D})(\mathbf{V}_1^*(\mathbf{U}_1\mathbf{D}))^{-1}\mathbf{V}_1^* = \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-1}\mathbf{V}_1^*.$$

Equations (7.10.41) and (7.10.42) follow from (5.9.11), p. 386. ∎

Formulas (7.10.40) and (7.10.42) are useful because all good matrix computation packages contain numerically stable SVD implementations from which $\mathbf{U}_1$ and $\mathbf{V}_1^*$ can be obtained. But, of course, the singular values are not needed in this application.

Example 7.10.8

Shell Game. As depicted in Figure 7.10.2, a pea is placed under one of four shells, and an agile manipulator quickly rearranges them by a sequence of discrete moves. At the end of each move the shell containing the pea has been shifted either to the left or right by only one position according to the following rules.

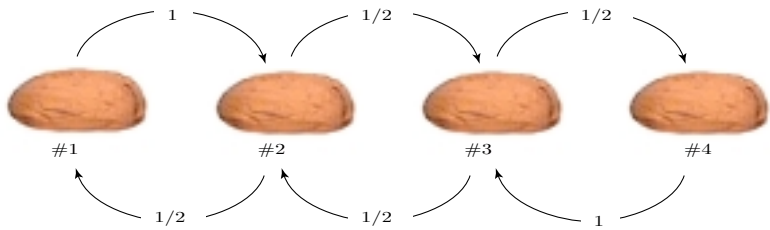


FIGURE 7.10.2

When the pea is under shell #1, it is moved to position #2, and if the pea is under shell #4, it is moved to position #3. When the pea is under shell #2 or #3, it is equally likely to be moved one position to the left or to the right.

Problem 1: Given that we know something about where the pea starts, what is the probability of finding the pea in any given position after k moves?

Problem 2: In the long run, what proportion of time does the pea occupy each of the four positions?

Solution to Problem 1: Let $p_j(k)$ denote the probability that the pea is in position j after the kth move, and translate the given information into four difference equations by writing

$$\begin{aligned} p_1(k) &= \frac{p_2(k-1)}{2} \\ p_2(k) &= p_1(k-1) + \frac{p_3(k-1)}{2} \\ p_3(k) &= \frac{p_2(k-1)}{2} + p_4(k-1) \\ p_4(k) &= \frac{p_3(k-1)}{2} \end{aligned} \quad \text{or} \quad \begin{pmatrix} p_1(k) \\ p_2(k) \\ p_3(k) \\ p_4(k) \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} p_1(k-1) \\ p_2(k-1) \\ p_3(k-1) \\ p_4(k-1) \end{pmatrix}.$$

The matrix equation on the right-hand side is a homogeneous difference equation $\mathbf{p}(k) = \mathbf{A}\mathbf{p}(k-1)$ whose solution, from (7.10.4), is $\mathbf{p}(k) = \mathbf{A}^k\mathbf{p}(0)$, and thus Problem 1 is solved. For example, if you know that the pea is initially under shell #2, then $\mathbf{p}(0) = \mathbf{e}_2$, and after six moves the probability that the pea is in the fourth position is $p_4(6) = \left[\mathbf{A}^6\mathbf{e}_2\right]_4 = 21/64$. If you don't know exactly where the pea starts, but you assume that it is equally likely to start under any one of the four shells, then $\mathbf{p}(0) = (1/4,\ 1/4,\ 1/4,\ 1/4)^T$, and the probabilities

for occupying the four positions after six moves are given by $\mathbf{p}(6) = \mathbf{A}^6\mathbf{p}(0)$, or

$$\begin{pmatrix} p_1(6) \\ p_2(6) \\ p_3(6) \\ p_4(6) \end{pmatrix} = \begin{pmatrix} 11/32 & 0 & 21/64 & 0 \\ 0 & 43/64 & 0 & 21/32 \\ 21/32 & 0 & 43/64 & 0 \\ 0 & 21/64 & 0 & 11/32 \end{pmatrix} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \frac{1}{256} \begin{pmatrix} 43 \\ 85 \\ 85 \\ 43 \end{pmatrix}.$$

Solution to Problem 2: There is a straightforward solution when $\mathbf{A}$ is a convergent matrix because if $\mathbf{A}^k \to \mathbf{G}$ as $k \to \infty$, then $\mathbf{p}(k) \to \mathbf{G}\mathbf{p}(0) = \mathbf{p}$, and the components in this limiting (or steady-state) vector $\mathbf{p}$ provide the answer. Intuitively, if $\mathbf{p}(k) \to \mathbf{p}$, then after awhile $\mathbf{p}(k)$ is practically constant, so the probability that the pea occupies a particular position remains essentially the same move after move. Consequently, the components in $\lim_{k\to\infty} \mathbf{p}(k)$ reveal the proportion of time spent in each position over the long run. For example, if $\lim_{k\to\infty} \mathbf{p}(k) = (1/6,\, 1/3,\, 1/3,\, 1/6)^T$, then, as the game runs on indefinitely, the pea is expected to be under shell #1 for about 16.7% of the time, under shell #2 for about 33.3% of the time, etc.

A Fly in the Ointment: Everything above rests on the assumption that $\mathbf{A}$ is convergent. But $\mathbf{A}$ is *not* convergent for the shell game because a bit of computation reveals that $\sigma(\mathbf{A}) = \{\pm 1,\, \pm(1/2)\}$. That is, there is an eigenvalue other than 1 on the unit circle, so (7.10.33) guarantees that $\lim_{k\to\infty} \mathbf{A}^k$ does not exist. Consequently, there's no limiting solution $\mathbf{p}$ to the difference equation $\mathbf{p}(k) = \mathbf{A}\mathbf{p}(k-1)$, and the intuitive analysis given above does not apply.

Cesàro to the Rescue: However, $\mathbf{A}$ is summable because $\rho(\mathbf{A}) = 1$, and every eigenvalue on the unit circle is semisimple—these are the conditions in (7.10.36). So as $k \to \infty$,

$$\left(\frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k}\right)\mathbf{p}(0) \to \mathbf{G}\mathbf{p}(0) = \mathbf{p}.$$

The job now is to interpret the meaning of this Cesàro limit in the context of the shell game. To do so, focus on a particular position—say the jth one—and set up "counting functions" (random variables) defined as

$$X(0) = \begin{cases} 1 & \text{if the pea starts under shell } j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X(i) = \begin{cases} 1 & \text{if the pea is under shell } j \text{ after the } i^{th} \text{ move,} \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, 2, 3, \ldots.$$

Notice that $X(0) + X(1) + \cdots + X(k-1)$ counts the number of times the pea occupies position j before the kth move, so $\big(X(0) + X(1) + \cdots + X(k-1)\big)/k$

represents the *fraction* of times that the pea is under shell j before the kth move. Since the expected (or mean) value of $X(i)$ is, by definition,

$$E[X(i)] = 1 \times P\big(X(i) = 1\big) \ + \ 0 \times P\big(X(i) = 0\big) = p_j(i),$$

and since expectation is linear ($E[\alpha X(i) + X(h)] = \alpha E[X(i)] + E[X(h)]$), the expected fraction of times that the pea occupies position j before move k is

$$\begin{aligned} E\left[\frac{X(0) + X(1) + \cdots + X(k-1)}{k}\right] &= \frac{E[X(0)] + E[X(1)] + \cdots + E[X(k-1)]}{k} \\ &= \frac{p_j(0) + p_j(1) + \cdots + p_j(k-1)}{k} = \left[\frac{\mathbf{p}(0) + \mathbf{p}(1) + \cdots + \mathbf{p}(k-1)}{k}\right]_j \\ &= \left[\frac{\mathbf{p}(0) + \mathbf{A}\mathbf{p}(0) + \cdots + \mathbf{A}^{k-1}\mathbf{p}(0)}{k}\right]_j = \left[\left(\frac{\mathbf{I} + \mathbf{A} + \cdots + \mathbf{A}^{k-1}}{k}\right)\mathbf{p}(0)\right]_j \\ &\to [\mathbf{G}\mathbf{p}(0)]_j \,. \end{aligned}$$

In other words, as the game progresses indefinitely, the components of the Cesàro limit $\mathbf{p} = \mathbf{G}\mathbf{p}(0)$ provide the expected proportion of times that the pea is under each shell, and this is exactly what we wanted to know.

Computing the Limiting Vector. Of course, $\mathbf{p}$ can be determined by first computing $\mathbf{G}$ with a full-rank factorization of $\mathbf{I} - \mathbf{A}$ as described in (7.10.41), but there is some special structure in this problem that can be exploited to make the task easier. Recall from (7.2.12) on p. 518 that if λ is a simple eigenvalue for $\mathbf{A}$, and if $\mathbf{x}$ and $\mathbf{y}^*$ are respective right-hand and left-hand eigenvectors associated with λ, then $\mathbf{x}\mathbf{y}^*/\mathbf{y}^*\mathbf{x}$ is the projector onto $N(\lambda\mathbf{I} - \mathbf{A})$ along $R(\lambda\mathbf{I} - \mathbf{A})$. We can use this because, for the shell game, $\lambda = 1$ is a simple eigenvalue for $\mathbf{A}$. Furthermore, we get an associated left-hand eigenvector for free—namely, $\mathbf{e}^T = (1, 1, 1, 1)$—because each column sum of $\mathbf{A}$ is one, so $\mathbf{e}^T\mathbf{A} = \mathbf{e}^T$. Consequently, if $\mathbf{x}$ is any right-hand eigenvector of $\mathbf{A}$ associated with $\lambda = 1$, then (by noting that $\mathbf{e}^T\mathbf{p}(0) = p_1(0) + p_2(0) + p_3(0) + p_4(0) = 1$) the limiting vector is given by

$$\mathbf{p} = \mathbf{G}\mathbf{p}(0) = \frac{\mathbf{x}\mathbf{e}^T\mathbf{p}(0)}{\mathbf{e}^T\mathbf{x}} = \frac{\mathbf{x}}{\mathbf{e}^T\mathbf{x}} = \frac{\mathbf{x}}{\sum x_i}. \tag{7.10.44}$$

In other words, the limiting vector is obtained by normalizing any nonzero solution of $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$ to make the components sum to one. Not only does (7.10.44) show how to compute the limiting proportions, it also shows that *the limiting proportions are independent of the initial values in* $\mathbf{p}(0)$. For example, a simple calculation reveals that $\mathbf{x} = (1, 2, 2, 1)^T$ is one solution of $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}$, so the vector of limiting proportions is $\mathbf{p} = (1/6, 1/3, 1/3, 1/6)^T$. Therefore, if many moves are made, then, regardless of where the pea starts, we expect the pea to end up under shell #1 in about 16.7% of the moves, under #2 for about

33.3% of the moves, under #3 for about 33.3% of the moves, and under shell #4 for about 16.7% of the moves.

Note: The shell game (and its analysis) is a typical example of *a random walk with reflecting barriers,* and these problems belong to a broader classification of stochastic processes known as *irreducible, periodic Markov chains.* (Markov chains are discussed in detail in §8.4 on p. 687.) The shell game is irreducible in the sense of Exercise 4.4.20 (p. 209), and it is periodic because the pea can return to given position only at definite periods, as reflected in the periodicity of the powers of $\mathbf{A}$. More details are given in Example 8.4.3 on p. 694.

Exercises for section 7.10

7.10.1. Which of the following are convergent, and which are summable?

$$\mathbf{A}=\begin{pmatrix} -1/2 & 3/2 & -3/2 \\ 1 & 0 & -1/2 \\ 1 & -1 & 1/2 \end{pmatrix}. \ \mathbf{B}=\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \ \mathbf{C}=\begin{pmatrix} -1 & -2 & -3/2 \\ 1 & 2 & 1 \\ 1 & 1 & 3/2 \end{pmatrix}.$$

7.10.2. For the matrices in Exercise 7.10.1, evaluate the limit of each convergent matrix, and evaluate the Cesàro limit for each summable matrix.

7.10.3. Verify that the expressions in (7.10.4) are indeed the solutions to the difference equations in (7.10.3).

7.10.4. Determine the limiting vector for the shell game in Example 7.10.8 by first computing the Cesàro limit $\mathbf{G}$ with a full-rank factorization.

7.10.5. Verify that the expressions in (7.10.4) are indeed the solutions to the difference equations in (7.10.3).

7.10.6. Prove that if there exists a matrix norm such that $\|\mathbf{A}\| < 1$, then $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$.

7.10.7. By examining the iteration matrix, compare the convergence of Jacobi's method and the Gauss–Seidel method for each of the following coefficient matrices with an arbitrary right-hand side. Explain why this shows that neither method can be universally favored over the other.

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}. \qquad \mathbf{A}_2 = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

7.10.8. Let $\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ (the finite-difference Example 1.4.1, p. 19).

(a) Verify that $\mathbf{A}$ satisfies the special case conditions given in Example 7.10.6 that guarantee the validity of (7.10.24).

(b) Determine the optimum SOR relaxation parameter.

(c) Find the asymptotic rates of convergence for Jacobi, Gauss–Seidel, and optimum SOR.

(d) Use $\mathbf{x}(0) = (1, 1, 1)^T$ and $\mathbf{b} = (2, 4, 6)^T$ to run through several steps of Jacobi, Gauss–Seidel, and optimum SOR to solve $\mathbf{Ax} = \mathbf{b}$ until you can see a convergence pattern.

7.10.9. Prove that if $\rho(\mathbf{H}_\omega) < 1$, where $\mathbf{H}_\omega$ is the iteration matrix for the SOR method, then $0 < \omega < 2$. **Hint:** Use $\det(\mathbf{H}_\omega)$ to show $|\lambda_k| \geq |1 - \omega|$ for some $\lambda_k \in \sigma(\mathbf{H}_\omega)$.

7.10.10. Show that the spectral radius of the Jacobi iteration matrix for the discrete Laplacian $\mathbf{L}_{n^2 \times n^2}$ described in Example 7.6.2 (p. 563) is $\rho(\mathbf{H}_J) = \cos \pi/(n+1)$.

7.10.11. Consider a scalar sequence $\{\alpha_1, \alpha_2, \alpha_3, \ldots\}$ and the associated Cesàro sequence of averages $\{\mu_1, \mu_2, \mu_3, \ldots\}$, where $\mu_n = (\alpha_1 + \alpha_2 + \cdots + \alpha_n)/n$. Prove that if $\{\alpha_n\}$ converges to α, then $\{\mu_n\}$ also converges to α.

Note: Like scalars, a vector sequence $\{\mathbf{v}_n\}$ in a finite-dimensional space converges to $\mathbf{v}$ if and only if for each $\epsilon > 0$ there is a natural number $N = N(\epsilon)$ such that $\|\mathbf{v}_n - \mathbf{v}\| < \epsilon$ for all $n \geq N$, and, by virtue of Example 5.1.3 (p. 276), it doesn't matter which norm is used. Therefore, your proof should also be valid for vectors (and matrices).

7.10.12. **M-matrices Revisited.** For matrices with nonpositive off-diagonal entries (Z-matrices), prove that the following statements are equivalent.

(a) $\mathbf{A}$ is an M-matrix.

(b) All *leading* principal minors of $\mathbf{A}$ are positive.

(c) $\mathbf{A}$ has an LU factorization, and both $\mathbf{L}$ and $\mathbf{U}$ are M-matrices.

(d) There exists a vector $\mathbf{x} > \mathbf{0}$ such that $\mathbf{Ax} > \mathbf{0}$.

(e) Each $a_{ii} > 0$ and $\mathbf{AD}$ is diagonally dominant for some diagonal matrix $\mathbf{D}$ with positive diagonal entries.

(f) $\mathbf{Ax} \geq \mathbf{0}$ implies $\mathbf{x} \geq \mathbf{0}$.

7.10.13. Index by Full-Rank Factorization. Suppose that $\lambda \in \sigma(\mathbf{A})$, and let $\mathbf{M}_1 = \mathbf{A} - \lambda\mathbf{I}$. The following procedure yields the value of $index(\lambda)$.

Factor $\mathbf{M}_1 = \mathbf{B}_1\mathbf{C}_1$ as a full-rank factorization.
Set $\mathbf{M}_2 = \mathbf{C}_1\mathbf{B}_1$.
Factor $\mathbf{M}_2 = \mathbf{B}_2\mathbf{C}_2$ as a full-rank factorization.
Set $\mathbf{M}_3 = \mathbf{C}_2\mathbf{B}_2$.
$\vdots$

In general, $\mathbf{M}_i = \mathbf{C}_{i-1}\mathbf{B}_{i-1}$, where $\mathbf{M}_{i-1} = \mathbf{B}_{i-1}\mathbf{C}_{i-1}$ is a full-rank factorization.

(a) Explain why this procedure must eventually produce a matrix $\mathbf{M}_k$ that is either nonsingular or zero.

(b) Prove that if k is the smallest positive integer such that $\mathbf{M}_k^{-1}$ exists or $\mathbf{M}_k = \mathbf{0}$, then

$$index(\lambda) = \begin{cases} k-1 & \text{if } \mathbf{M}_k \text{ is nonsingular,} \\ k & \text{if } \mathbf{M}_k = \mathbf{0}. \end{cases}$$

7.10.14. Use the procedure in Exercise 7.10.13 to find the index of each eigenvalue of $\mathbf{A} = \begin{pmatrix} -3 & -8 & -9 \\ 5 & 11 & 9 \\ -1 & -2 & 1 \end{pmatrix}$. **Hint:** $\sigma(\mathbf{A}) = \{4, 1\}$.

7.10.15. Let $\mathbf{A}$ be the matrix given in Exercise 7.10.14.

(a) Find the Jordan form for $\mathbf{A}$.

(b) For any function f defined at $\mathbf{A}$, find the Hermite interpolation polynomial that is described in Example 7.9.4 (p. 606), and describe $f(\mathbf{A})$.

7.10.16. Limits and Group Inversion. Given a matrix $\mathbf{B}_{n\times n}$ of rank r such that $index(\mathbf{B}) \le 1$ (i.e., $index(\lambda = 0) \le 1$), the Jordan form for $\mathbf{B}$ looks like $\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{r\times r} \end{pmatrix} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$, so $\mathbf{B} = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}\mathbf{P}^{-1}$, where $\mathbf{C}$ is nonsingular. This implies that $\mathbf{B}$ belongs to an algebraic group $\mathcal{G}$ with respect to matrix multiplication, and the inverse of $\mathbf{B}$ in $\mathcal{G}$ is $\mathbf{B}^{\#} = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix}\mathbf{P}^{-1}$. Naturally, $\mathbf{B}^{\#}$ is called the ***group inverse*** of $\mathbf{B}$. The group inverse is a special case of the Drazin inverse discussed in Example 5.10.5 on p. 399, and properties of group inversion are developed in Exercises 5.10.11–5.10.13 on p. 402. Prove that if $\lim_{k\to\infty} \mathbf{A}^k$ exists, and if $\mathbf{B} = \mathbf{I} - \mathbf{A}$, then

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{I} - \mathbf{B}\mathbf{B}^{\#}.$$

In other words, the limiting matrix can be characterized as the difference of two identity elements—$\mathbf{I}$ is the identity in the multiplicative group of nonsingular matrices, and $\mathbf{B}\mathbf{B}^{\#}$ is the identity element in the multiplicative group containing $\mathbf{B}$.

7.10.17. If $\mathbf{M}_{n\times n}$ is a group matrix (i.e., if $index(\mathbf{M}) \leq 1$), then the group inverse of $\mathbf{M}$ can be characterized as the unique solution $\mathbf{M}^{\#}$ of the equations $\mathbf{M}\mathbf{M}^{\#}\mathbf{M} = \mathbf{M}$, $\mathbf{M}^{\#}\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}$, and $\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}\mathbf{M}$. In fact, some authors use these equations to define $\mathbf{M}^{\#}$. Use this characterization to show that if $\mathbf{M} = \mathbf{B}\mathbf{C}$ is any full-rank factorization of $\mathbf{M}$, then $\mathbf{M}^{\#} = \mathbf{B}(\mathbf{C}\mathbf{B})^{-2}\mathbf{C}$. In particular, if $\mathbf{M} = \mathbf{U}_1\mathbf{D}\mathbf{V}_1^*$ is the full-rank factorization derived from the singular value decomposition as described in (7.10.38), then

$$\begin{aligned}\mathbf{M}^{\#} &= \mathbf{U}_1\mathbf{D}^{-1/2}(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{D}^{-1/2}\mathbf{V}_1^* \\ &= \mathbf{U}_1\mathbf{D}^{-1}(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{V}_1^* \\ &= \mathbf{U}_1(\mathbf{V}_1^*\mathbf{U}_1)^{-2}\mathbf{D}^{-1}\mathbf{V}_1^*.\end{aligned}$$

7.11 MINIMUM POLYNOMIALS AND KRYLOV METHODS

The characteristic polynomial plays a central role in the theoretical development of linear algebra and matrix analysis, but it is not alone in this respect. There are other polynomials that occur naturally, and the purpose of this section is to explore some of them.

In this section it is convenient to consider the characteristic polynomial of $\mathbf{A} \in \mathcal{C}^{n\times n}$ to be $c(x) = \det(x\mathbf{I} - \mathbf{A})$. This differs from the definition given on p. 492 only in the sense that the coefficients of $c(x) = \det(x\mathbf{I} - \mathbf{A})$ have different signs than the coefficients of $\hat{c}(x) = \det(\mathbf{A} - x\mathbf{I})$. In particular, $c(x)$ is a *monic polynomial* (i.e., its leading coefficient is 1), whereas the leading coefficient of $\hat{c}(x)$ is $(-1)^n$. (Of course, the roots of c and $\hat{c}$ are identical.)

Monic polynomials $p(x)$ such that $p(\mathbf{A}) = \mathbf{0}$ are said to be ***annihilating polynomials*** for $\mathbf{A}$. For example, the Cayley–Hamilton theorem (pp. 509, 532) guarantees that $c(x)$ is an annihilating polynomial of degree n.

Minimum Polynomial for a Matrix

There is a unique annihilating polynomial for $\mathbf{A}$ of minimal degree, and this polynomial, denoted by $m(x)$, is called the ***minimum polynomial*** for $\mathbf{A}$. The Cayley–Hamilton theorem guarantees that $\deg[m(x)] \leq n$.

Proof. Only uniqueness needs to be proven. Let k be the smallest degree of any annihilating polynomial for $\mathbf{A}$. There is a unique annihilating polynomial for $\mathbf{A}$ of degree k because if there were two different annihilating polynomials $p_1(x)$ and $p_2(x)$ of degree k, then $d(x) = p_1(x) - p_2(x)$ would be a nonzero polynomial such that $d(\mathbf{A}) = \mathbf{0}$ and $\deg[d(x)] < k$. Dividing $d(x)$ by its leading coefficient would produce an annihilating polynomial of degree less than k, the minimal degree, and this is impossible. ∎

The first problem is to describe what the minimum polynomial $m(x)$ for $\mathbf{A} \in \mathcal{C}^{n\times n}$ looks like, and the second problem is to uncover the relationship between $m(x)$ and the characteristic polynomial $c(x)$. The Jordan form for $\mathbf{A}$ reveals everything. Suppose that $\mathbf{A} = \mathbf{PJP}^{-1}$, where $\mathbf{J}$ is in Jordan form. Since $p(\mathbf{A}) = \mathbf{0}$ if and only if $p(\mathbf{J}) = \mathbf{0}$ or, equivalently, $p(\mathbf{J}_\star) = \mathbf{0}$ for each Jordan block $\mathbf{J}_\star$, it's clear that $m(x)$ is the monic polynomial of smallest degree that annihilates all Jordan blocks. If $\mathbf{J}_\star$ is a $k \times k$ Jordan block associated with an eigenvalue λ, then (7.9.2) on p. 600 insures that $p(\mathbf{J}_\star) = \mathbf{0}$ if and only if $p^{(i)}(\lambda) = 0$ for $i = 0, 2, \ldots, k-1$, and this happens if and only if $p(x) = (x-\lambda)^k q(x)$ for some polynomial $q(x)$. Since this must be true for all Jordan blocks associated with λ, it must be true for the *largest* Jordan block associated with λ, and thus the minimum degree monic polynomial that

annihilates all Jordan blocks associated with λ is

$$p_\lambda(x) = (x-\lambda)^{k_\lambda}, \quad \text{where} \quad k_\lambda = index(\lambda).$$

Since the minimum polynomial for $\mathbf{A}$ must annihilate the largest Jordan block associated with each $\lambda_j \in \sigma(\mathbf{A})$, it follows that

$$m(x) = (x-\lambda_1)^{k_1}(x-\lambda_2)^{k_2}\cdots(x-\lambda_s)^{k_s}, \quad \text{where} \quad k_j = index(\lambda_j) \tag{7.11.1}$$

is the minimum polynomial for $\mathbf{A}$.

Example 7.11.1

Minimum Polynomial, Gram–Schmidt, and QR. If you are willing to compute the eigenvalues λ_j and their indicies k_j for a given $\mathbf{A} \in \mathcal{C}^{n\times n}$, then, as shown in (7.11.1), the minimum polynomial for $\mathbf{A} \in \mathcal{C}^{n\times n}$ is obtained by setting $m(x) = (x-\lambda_1)^{k_1}(x-\lambda_2)^{k_2}\cdots(x-\lambda_s)^{k_s}$. But finding the eigenvalues and their indicies can be a substantial task, so let's consider how we might construct $m(x)$ without computing eigenvalues. An approach based on first principles is to determine the first matrix $\mathbf{A}^k$ for which $\{\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^k\}$ is linearly dependent. In other words, if k is the smallest positive integer such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$, then the minimum polynomial for $\mathbf{A}$ is

$$m(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j.$$

The Gram–Schmidt orthogonalization procedure (p. 309) with the standard inner product $\langle \mathbf{A}|\mathbf{B}\rangle = trace\,(\mathbf{A}^*\mathbf{B})$ (p. 286) is the perfect theoretical tool for determining k and the α_j's. Gram–Schmidt applied to $\{\mathbf{I}, \mathbf{A}, \mathbf{A}^2, \ldots\}$ begins by setting $\mathbf{U}_0 = \mathbf{I}/\|\mathbf{I}\|_F = \mathbf{I}/\sqrt{n}$, and it proceeds by sequentially computing

$$\mathbf{U}_j = \frac{\mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i}{\|\mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i\|_F} \quad \text{for} \quad j = 1, 2, \ldots \tag{7.11.2}$$

until $\mathbf{A}^k - \sum_{i=0}^{k-1} \langle \mathbf{U}_i | \mathbf{A}^k \rangle \mathbf{U}_i = \mathbf{0}$. The first such k is the smallest positive integer such that $\mathbf{A}^k \in span\,\{\mathbf{U}_0, \mathbf{U}_1, \ldots, \mathbf{U}_{k-1}\} = span\,\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\}$. The coefficients α_j such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$ are easily determined from the upper-triangular matrix $\mathbf{R}$ in the QR factorization produced by the Gram–Schmidt process. To see how, extend the notation in the discussion on p. 311 in an obvious way to write (7.11.2) in matrix form as

$$\left[\mathbf{I}\,|\,\mathbf{A}\,|\cdots|\,\mathbf{A}^k\right] = \left[\mathbf{U}_0\,|\,\mathbf{U}_1\,|\cdots|\,\mathbf{U}_k\right] \begin{pmatrix} \nu_0 & r_{01} & \cdots & r_{0k-1} & r_{0k} \\ 0 & \nu_1 & \cdots & r_{1k-1} & r_{1k} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & \nu_{k-1} & r_{k-1k} \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{7.11.3}$$

where $\nu_0 = \|\mathbf{I}\|_F = \sqrt{n}$, $\nu_j = \left\|\mathbf{A}^j - \sum_{i=0}^{j-1} \langle \mathbf{U}_i | \mathbf{A}^j \rangle \mathbf{U}_i\right\|_F$, and $r_{ij} = \langle \mathbf{U}_i | \mathbf{A}^j \rangle$.

If we set $\mathbf{R} = \begin{pmatrix} \nu_0 & \cdots & r_{0k-1} \\ & \ddots & \vdots \\ & & \nu_{k-1} \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} r_{0k} \\ \vdots \\ r_{k-1k} \end{pmatrix}$, then (7.11.3) implies that

$\mathbf{A}^k = \left[\mathbf{U}_0 | \cdots | \mathbf{U}_{k-1}\right]\mathbf{c} = \left[\mathbf{I} | \cdots | \mathbf{A}^{k-1}\right]\mathbf{R}^{-1}\mathbf{c}$, so $\mathbf{R}^{-1}\mathbf{c} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{k-1} \end{pmatrix}$ contains the coefficients such that $\mathbf{A}^k = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j$, and thus the coefficients in the minimum polynomial are determined.

Caution! While Gram–Schmidt works fine to produce $m(x)$ in exact arithmetic, things are not so nice in floating-point arithmetic. For example, if $\mathbf{A}$ has a dominant eigenvalue, then, as explained in the power method (Example 7.3.7, p. 533), $\mathbf{A}^k$ asymptotically approaches the dominant spectral projector $\mathbf{G}_1$, so, as k grows, $\mathbf{A}^k$ becomes increasingly close to $span\left\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\right\}$. Consequently, finding the first $\mathbf{A}^k$ that is truly in $span\left\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{k-1}\right\}$ is an ill-conditioned problem, and Gram–Schmidt may not work well in floating-point arithmetic—the modified Gram–Schmidt algorithm (p. 316), or a version of Householder reduction (p. 341), or Arnoldi's method (p. 653) works better. Fortunately, explicit knowledge of the minimum polynomial often is not needed in applied work.

The relationship between the characteristic polynomial $c(x)$ and the minimum polynomial $m(x)$ for $\mathbf{A}$ is now transparent. Since

$$c(x) = (x-\lambda_1)^{a_1}(x-\lambda_2)^{a_2}\cdots(x-\lambda_s)^{a_s}, \quad \text{where} \quad a_j = alg\ mult\,(\lambda_j),$$

and

$$m(x) = (x-\lambda_1)^{k_1}(x-\lambda_2)^{k_2}\cdots(x-\lambda_s)^{k_s}, \quad \text{where} \quad k_j = index(\lambda_j),$$

it's clear that $m(x)$ divides $c(x)$. Furthermore, $m(x) = c(x)$ if and only if $alg\ mult\,(\lambda_j) = index(\lambda_j)$ for each $\lambda_j \in \sigma(\mathbf{A})$. Matrices for which $m(x) = c(x)$ are said to be ***nonderogatory matrices,*** and they are precisely the ones for which $geo\ mult\,(\lambda_j) = 1$ for each eigenvalue λ_j because

$$\begin{aligned} m(x) = c(x) &\iff alg\ mult\,(\lambda_j) = index(\lambda_j) \text{ for each } j \\ &\iff \text{there is only one Jordan block for each } \lambda_j \\ &\iff \text{there is only one independent eigenvector for each } \lambda_j \\ &\iff geo\ mult\,(\lambda_j) = 1 \text{ for each } \lambda_j. \end{aligned}$$

In addition to dividing the characteristic polynomial $c(x)$, the minimum polynomial $m(x)$ divides all other annihilating polynomials $p(x)$ for $\mathbf{A}$ because $\deg[m(x)] \le \deg[p(x)]$ insures the existence of polynomials $q(x)$ and $r(x)$ (quotient and remainder) such that

$$p(x) = m(x)q(x) + r(x), \quad \text{where} \quad \deg[r(x)] < \deg[m(x)].$$

Since

$$\mathbf{0} = p(\mathbf{A}) = m(\mathbf{A})q(\mathbf{A}) + r(\mathbf{A}) = r(\mathbf{A}),$$

it follows that $r(x) = 0$; otherwise $r(x)$, when normalized to be monic, would be an annihilating polynomial having degree smaller than the degree of the minimum polynomial.

The structure of the minimum polynomial for $\mathbf{A}$ is related to the diagonalizability of $\mathbf{A}$. By combining the fact that $k_j = index(\lambda_j)$ is the size of the largest Jordan block for λ_j with the fact that $\mathbf{A}$ is diagonalizable if and only if all Jordan blocks are 1×1, it follows that $\mathbf{A}$ is diagonalizable if and only if $k_j = 1$ for each j, which, by (7.11.1), is equivalent to saying that $m(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_s)$. In other words, $\mathbf{A}$ is diagonalizable if and only if its minimum polynomial is the product of distinct linear factors.

Below is a summary of the preceding observations about properties of $m(x)$.

Properties of the Minimum Polynomial

Let $\mathbf{A} \in \mathcal{C}^{n\times n}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$.

- The minimum polynomial of $\mathbf{A}$ is the unique monic polynomial $m(x)$ of minimal degree such that $m(\mathbf{A}) = \mathbf{0}$.
- $m(x) = (x - \lambda_1)^{k_1}(x - \lambda_2)^{k_2} \cdots (x - \lambda_s)^{k_s}$, where $k_j = index(\lambda_j)$.
- $m(x)$ divides every polynomial $p(x)$ such that $p(\mathbf{A}) = \mathbf{0}$. In particular, $m(x)$ divides the characteristic polynomial $c(x)$. (7.11.4)
- $m(x) = c(x)$ if and only if $geo\ mult(\lambda_j) = 1$ for each λ_j or, equivalently, $alg\ mult(\lambda_j) = index(\lambda_j)$ for each j, in which case $\mathbf{A}$ is called a ***nonderogatory matrix.***
- $\mathbf{A}$ is diagonalizable if and only if $m(x) = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_s)$ (i.e., if and only if $m(x)$ is a product of distinct linear factors).

The next immediate aim is to extend the concept of the minimum polynomial for a matrix to formulate the notion of a minimum polynomial for a vector. To do so, it's helpful to introduce Krylov[86] sequences, subspaces, and matrices.

[86] Aleksei Nikolaevich Krylov (1863–1945) showed in 1931 how to use sequences of the form $\{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}, \ldots\}$ to construct the characteristic polynomial of a matrix (see Example 7.11.3 on p. 649). Krylov was a Russian applied mathematician whose scientific interests arose from his early training in naval science that involved the theories of buoyancy, stability, rolling and pitching, vibrations, and compass theories. Krylov served as the director of the Physics–Mathematics Institute of the Soviet Academy of Sciences from 1927 until 1932, and in 1943 he was awarded a "state prize" for his work on compass theory. Krylov was made a "hero of

Krylov Sequences, Subspaces, and Matrices

For $\mathbf{A} \in \mathcal{C}^{n\times n}$ and $\mathbf{0} \neq \mathbf{b} \in \mathcal{C}^{n\times 1}$, we adopt the following terminology.

- $\{\mathbf{b},\, \mathbf{Ab},\, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{j-1}\mathbf{b}\}$ is called a ***Krylov sequence.***
- $\mathcal{K}_j = span\left\{\mathbf{b},\, \mathbf{Ab},\, \ldots, \mathbf{A}^{j-1}\mathbf{b}\right\}$ is called a ***Krylov subspace.***
- $\mathbf{K}_{n\times j} = \left(\mathbf{b}\,|\,\mathbf{Ab}\,|\cdots|\,\mathbf{A}^{j-1}\mathbf{b}\right)$ is called a ***Krylov matrix.***

Since $\dim(\mathcal{K}_j) \leq n$ (because $\mathcal{K}_j \subseteq \mathcal{C}^{n\times 1}$), there is a first vector $\mathbf{A}^k\mathbf{b}$ in the Krylov sequence that is a linear combination of preceding Krylov vectors. If

$$\mathbf{A}^k\mathbf{b} = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j\mathbf{b}, \quad \text{then we define} \quad v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j,$$

and we say that $v(x)$ is an ***annihilating polynomial for*** $\mathbf{b}$ ***relative to*** $\mathbf{A}$ because $v(x)$ is a monic polynomial such that $v(\mathbf{A})\mathbf{b} = \mathbf{0}$. The argument on p. 642 that establishes uniqueness of the minimum polynomial for matrices can be reapplied to prove that for each matrix–vector pair $(\mathbf{A}, \mathbf{b})$ there is a unique annihilating polynomial of $\mathbf{b}$ relative to $\mathbf{A}$ that has minimal degree. These observations are formalized below.

Minimum Polynomial for a Vector

- The ***minimum polynomial for*** $\mathbf{b} \in \mathcal{C}^{n\times 1}$ ***relative to*** $\mathbf{A} \in \mathcal{C}^{n\times n}$ is defined to be the monic polynomial $v(x)$ of minimal degree such that $v(\mathbf{A})\mathbf{b} = \mathbf{0}$.
- If $\mathbf{A}^k\mathbf{b}$ is the first vector in the Krylov sequence $\{\mathbf{b},\, \mathbf{Ab},\, \mathbf{A}^3\mathbf{b}, \ldots\}$ that is a linear combination of preceding Krylov vectors (say $\mathbf{A}^k\mathbf{b} = \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j\mathbf{b}$), then $v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ (or $v(x) = 1$ when $\mathbf{b} = \mathbf{0}$) is the minimum polynomial for $\mathbf{b}$ relative to $\mathbf{A}$.

socialist labor," and he is one of a few mathematicians to have a lunar feature named in his honor—on the moon there is the "Crater Krylov."

So is the minimum polynomial for a matrix related to minimum polynomials for vectors? It seems intuitive that knowing the minimum polynomial of $\mathbf{b}$ relative to $\mathbf{A}$ for enough different vectors $\mathbf{b}$ should somehow lead to the minimum polynomial for $\mathbf{A}$. This is indeed the case, and here is how it's done. Recall that the least common multiple (LCM) of polynomials $v_1(x), \ldots, v_n(x)$ is the unique monic polynomial $l(x)$ such that

(i) each $v_i(x)$ divides $l(x)$;

(ii) if each $v_i(x)$ also divides $q(x)$, then $l(x)$ divides $q(x)$.

Minimum Polynomial as LCM

Let $\mathbf{A} \in \mathcal{C}^{n\times n}$, and let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be any basis for $\mathcal{C}^{n\times 1}$. If $v_i(x)$ is the minimum polynomial for $\mathbf{b}_i$ relative to $\mathbf{A}$, then the minimum polynomial $m(x)$ for $\mathbf{A}$ is the least common multiple of $v_1(x)$, $v_2(x)$, $\ldots$, $v_n(x)$. (7.11.5)

Proof. The strategy first is to prove that if $l(x)$ is the LCM of the $v_i(x)$'s, then $m(x)$ divides $l(x)$. Then prove the reverse by showing that $l(x)$ also divides $m(x)$. Since each $v_i(x)$ divides $l(x)$, it follows that $l(\mathbf{A})\mathbf{b}_i = \mathbf{0}$ for each i. In other words, $\mathcal{B} \subset N(l(\mathbf{A}))$, so $\dim N(l(\mathbf{A})) = n$ or, equivalently, $l(\mathbf{A}) = \mathbf{0}$. Therefore, by property (7.11.4) on p. 645, $m(x)$ divides $l(x)$. Now show that $l(x)$ divides $m(x)$. Since $m(\mathbf{A})\mathbf{b}_i = \mathbf{0}$ for every $\mathbf{b}_i$, it follows that $\deg[v_i(x)] < \deg[m(x)]$ for each i, and hence there exist polynomials $q_i(x)$ and $r_i(x)$ such that $m(x) = q_i(x)v_i(x) + r_i(x)$, where $\deg[r_i(x)] < \deg[v_i(x)]$. But

$$\mathbf{0} = m(\mathbf{A})\mathbf{b}_i = q_i(\mathbf{A})v_i(\mathbf{A})\mathbf{b}_i + r_i(\mathbf{A})\mathbf{b}_i = r_i(\mathbf{A})\mathbf{b}_i$$

insures $r_i(x) = 0$, for otherwise $r_i(x)$ (when normalized to be monic) would be an annihilating polynomial for $\mathbf{b}_i$ of degree smaller than the minimum polynomial for $\mathbf{b}_i$, which is impossible. In other words, each $v_i(x)$ divides $m(x)$, and this implies $l(x)$ must also divide $m(x)$. Therefore, since $m(x)$ and $l(x)$ are divisors of each other, it must be the case that $m(x) = l(x)$. ∎

The utility of this result is illustrated in the following development. We already know that associated with $n \times n$ matrix $\mathbf{A}$ is an nth-degree monic polynomial—namely, the characteristic polynomial $c(x) = \det(x\mathbf{I} - \mathbf{A})$. But the reverse is also true. That is, every n^{th}-degree monic polynomial is the characteristic polynomial of some $n \times n$ matrix.

Companion Matrix of a Polynomial

For each monic polynomial $p(x) = x^n + \alpha_{n-1}x^{n-1} + \cdots + \alpha_1 x + \alpha_0$, the ***companion matrix*** of $p(x)$ is defined (by G. Frobenius) to be

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\alpha_0 \\ 1 & 0 & \cdots & 0 & -\alpha_1 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 1 & 0 & -\alpha_{n-2} \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} \end{pmatrix}_{n\times n}. \tag{7.11.6}$$

- The polynomial $p(x)$ is both the characteristic and minimum polynomial for $\mathbf{C}$ (i.e., $\mathbf{C}$ is nonderogatory).

Proof. To prove that $\det(x\mathbf{I} - \mathbf{C}) = p(x)$, write $\mathbf{C} = \mathbf{N} - \mathbf{c}\mathbf{e}_n^T$, where

$$\mathbf{N} = \begin{pmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{pmatrix},$$

and use (6.2.3) on p. 475 to conclude that

$$\begin{aligned} \det(x\mathbf{I} - \mathbf{C}) &= \det(x\mathbf{I} - \mathbf{N})(1 + \mathbf{e}_n^T \det(x\mathbf{I} - \mathbf{N})^{-1}\mathbf{c}) \\ &= x^n \left(1 + \mathbf{e}_n^T \left(\frac{\mathbf{I}}{x} + \frac{\mathbf{N}}{x^2} + \frac{\mathbf{N}^2}{x^3} + \cdots + \frac{\mathbf{N}^{n-1}}{x^n}\right)\mathbf{c}\right) \\ &= x^n + \alpha_{n-1}x^{n-1} + \alpha_{n-2}x^{n-2} + \cdots + \alpha_0 \\ &= p(x). \end{aligned}$$

The fact that $p(x)$ is also the minimum polynomial for $\mathbf{C}$ is a consequence of (7.11.5). Set $\mathcal{B} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, and let $v_i(x)$ be the minimum polynomial of $\mathbf{e}_i$ with respect to $\mathbf{C}$. Observe that $v_1(x) = p(x)$ because $\mathbf{C}\mathbf{e}_j = \mathbf{e}_{j+1}$ for $j = 1, \ldots, n-1$, so

$$\{\mathbf{e}_1, \mathbf{C}\mathbf{e}_1, \mathbf{C}^2\mathbf{e}_1, \ldots, \mathbf{C}^{n-1}\mathbf{e}_1\} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_n\}$$

and

$$\mathbf{C}^n\mathbf{e}_1 = \mathbf{C}\mathbf{e}_n = \mathbf{C}_{*n} = -\sum_{j=0}^{n-1} \alpha_j \mathbf{e}_{j+1} = -\sum_{j=0}^{n-1} \alpha_j \mathbf{C}^j \mathbf{e}_1 \implies v_1(x) = p(x).$$

Since $v_1(x)$ divides the LCM of all $v_i(x)$'s (which we know from (7.11.5) to be the minimum polynomial $m(x)$ for $\mathbf{C}$), we conclude that $p(x)$ divides $m(x)$. But $m(x)$ always divides $p(x)$—recall (7.11.4)—so $m(x) = p(x)$. ∎

Example 7.11.2

Poor Man's Root Finder. The companion matrix is the source of what is often called the *poor man's root finder* because any general purpose algorithm designed to compute eigenvalues (e.g., the QR iteration on p. 535) can be applied to the companion matrix for a polynomial $p(x)$ to compute the roots of $p(x)$. When used in conjunction with (7.1.12) on p. 497, the companion matrix is also a *poor man's root bounder*. For example, it follows that if λ is a root of $p(x)$, then

$$|\lambda| \leq \|\mathbf{C}\|_\infty = \max\{|\alpha_0|,\, 1+|\alpha_1|, \ldots, 1+|\alpha_{n-1}|\} \leq 1 + \max|\alpha_i|.$$

The results on p. 647 insure that the minimum polynomial $v(x)$ for every nonzero vector $\mathbf{b}$ relative to $\mathbf{A} \in \mathcal{C}^{n\times n}$ divides the minimum polynomial $m(x)$ for $\mathbf{A}$, which in turn divides the characteristic polynomial $c(x)$ for $\mathbf{A}$, so it follows that every $v(x)$ divides $c(x)$. This suggests that it might be possible to construct $c(x)$ as a product of $v_i(x)$'s. In fact, this is what Krylov did in 1931, and the following example shows how he did it.

Example 7.11.3

Krylov's method for constructing the characteristic polynomial for $\mathbf{A} \in \mathcal{C}^{n\times n}$ as a product of minimum polynomials for vectors is as follows.

Starting with any nonzero vector $\mathbf{b}_{n\times 1}$, let $v_1(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ be the minimum polynomial for $\mathbf{b}$ relative to $\mathbf{A}$, and let $\mathbf{K}_1 = \big(\mathbf{b}\,|\,\mathbf{A}\mathbf{b}\,|\cdots|\,\mathbf{A}^{k-1}\mathbf{b}\big)_{n\times k}$ be the associated Krylov matrix. Notice that $rank\,(\mathbf{K}_1) = k$ (by definition of the minimum polynomial for $\mathbf{b}$). If $\mathbf{C}_1$ is the $k\times k$ companion matrix of $v(x)$ as described in (7.11.6), then direct multiplication shows that

$$\mathbf{K}_1\mathbf{C}_1 = \mathbf{A}\mathbf{K}_1. \tag{7.11.7}$$

If $k = n$, then $\mathbf{K}_1^{-1}\mathbf{A}\mathbf{K}_1 = \mathbf{C}_1$, so $v_1(x)$ must be the characteristic polynomial for $\mathbf{A}$, and there is nothing more to do. If $k < n$, then use any $n\times(n-k)$ matrix $\widetilde{\mathbf{K}}_1$ such that $\mathbf{K}_2 = \big(\mathbf{K}_1\,|\,\widetilde{\mathbf{K}}_1\big)_{n\times n}$ is nonsingular, and use (7.11.7) to write

$$\mathbf{A}\mathbf{K}_2 = \big(\mathbf{A}\mathbf{K}_1\,|\,\mathbf{A}\widetilde{\mathbf{K}}_1\big) = \big(\mathbf{K}_1\,|\,\widetilde{\mathbf{K}}_1\big)\begin{pmatrix}\mathbf{C}_1 & \mathbf{X}\\ \mathbf{0} & \mathbf{A}_2\end{pmatrix}, \quad \text{where} \quad \begin{pmatrix}\mathbf{X}\\ \mathbf{A}_2\end{pmatrix} = \mathbf{K}_2^{-1}\mathbf{A}\widetilde{\mathbf{K}}_1.$$

Therefore, $\mathbf{K}_2^{-1}\mathbf{A}\mathbf{K}_2 = \begin{pmatrix}\mathbf{C}_1 & \mathbf{X}\\ \mathbf{0} & \mathbf{A}_2\end{pmatrix}$, and hence

$$c(x) = \det(x\mathbf{I} - \mathbf{A}) = \det(x\mathbf{I} - \mathbf{C}_1)\det(x\mathbf{I} - \mathbf{A}_2) = v_1(x)\det(x\mathbf{I} - \mathbf{A}_2).$$

Repeat the process on $\mathbf{A}_2$. If the Krylov matrix on the second time around is nonsingular, then $c(x) = v_1(x)v_2(x)$; otherwise $c(x) = v_1(x)v_2(x)\det(x\mathbf{I} - \mathbf{A}_3)$ for some matrix $\mathbf{A}_3$. Continuing in this manner until a nonsingular Krylov matrix is obtained—say at the mth step—produces a nonsingular matrix $\mathbf{K}$ such that

$$\mathbf{K}^{-1}\mathbf{A}\mathbf{K} = \begin{pmatrix} \mathbf{C}_1 & \cdots & \star \\ & \ddots & \vdots \\ & & \mathbf{C}_m \end{pmatrix} = \mathbf{H}, \tag{7.11.8}$$

where the $\mathbf{C}_j$'s are companion matrices, and thus $c(x) = v_1(x)v_2(x)\cdots v_m(x)$.

Note: All companion matrices are upper-Hessenberg matrices as described in Example 5.7.4 (p. 350)—e.g., a 5×5 Hessenberg form is

$$\mathbf{H}_5 = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

Since the matrix $\mathbf{H}$ in (7.11.8) is upper Hessenberg, we see that Krylov's method boils down to a recipe for using Krylov sequences to build a similarity transformation that will reduce $\mathbf{A}$ to upper-Hessenberg form. In effect, this means that most information about $\mathbf{A}$ can be derived from Krylov sequences and the associated Hessenberg form $\mathbf{H}$. This is the real message of this example.

Deriving information about $\mathbf{A}$ by using a Hessenberg form and a Krylov similarity transformation as shown in (7.11.8) has some theoretical appeal, but it's not a practical idea as far as computation is concerned. Krylov sequences tend to be nearly linearly dependent sets because, as the power method of Example 7.3.7 (p. 533) indicates, the directions of the vectors $\mathbf{A}^k\mathbf{b}$ want to converge to the direction of an eigenvector for $\mathbf{A}$, so, as k grows, the vectors in a Krylov sequence become ever closer to being multiples of each other. This means that Krylov matrices tend to be ill conditioned. Putting conditioning issues aside, there is still a problem with computational efficiency because $\mathbf{K}$ is usually a dense matrix (one with a preponderance of nonzero entries) even when $\mathbf{A}$ is sparse (which it often is in applied work), so the amount of arithmetic involved in the reduction (7.11.8) is prohibitive.

However, these objections often can be overcome by replacing a Krylov matrix $\mathbf{K} = \big(\mathbf{b} \,|\, \mathbf{A}\mathbf{b} \,|\, \cdots \,|\, \mathbf{A}^{k-1}\mathbf{b}\big)$ with its QR factorization $\mathbf{K} = \mathbf{Q}_{n\times k}\mathbf{R}_{k\times k}$. Doing so in (7.11.7) (and dropping the subscript) produces

$$\mathbf{A}\mathbf{K} = \mathbf{K}\mathbf{C} \implies \mathbf{A}\mathbf{Q}\mathbf{R} = \mathbf{Q}\mathbf{R}\mathbf{C} \implies \mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{R}\mathbf{C}\mathbf{R}^{-1} = \mathbf{H}. \tag{7.11.9}$$

While $\mathbf{H} = \mathbf{R}\mathbf{C}\mathbf{R}^{-1}$ is no longer a companion matrix, it's still in upper-Hessenberg form (convince yourself by writing out the pattern for the 4×4 case). In other words, an orthonormal basis for a Krylov subspace can reduce a

matrix to upper-Hessenberg form. Since matrices with orthonormal columns are perfectly conditioned, the first objection raised above is overcome. The second objection concerning computational efficiency is dealt with in Examples 7.11.4 and 7.11.5.

If $k < n$, then $\mathbf{Q}$ is not square, and $\mathbf{Q}^*\mathbf{A}\mathbf{Q} = \mathbf{H}$ is not a similarity transformation, so it would be wrong to conclude that $\mathbf{A}$ and $\mathbf{H}$ have the same spectral properties. Nevertheless, it's often the case that the eigenvalues of $\mathbf{H}$, which are called the ***Ritz values*** for $\mathbf{A}$, are remarkably good approximations to the extreme eigenvalues of $\mathbf{A}$, especially when $\mathbf{A}$ is hermitian. This is somewhat intuitive because $\mathbf{Q}^*\mathbf{A}\mathbf{Q}$ can be viewed as a generalization of (7.5.4) on p. 549 that says $\lambda_{\max} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x}$ and $\lambda_{\min} = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}\mathbf{x}$. The results of Exercise 5.9.15 (p. 392) can be used to argue the point further.

Example 7.11.4

Lanczos[87] Tridiagonalization Algorithm. The fact that the matrix $\mathbf{H}$ in (7.11.9) is upper Hessenberg is particularly nice when $\mathbf{A}$ is real and symmetric because $\mathbf{A}^T = \mathbf{A}$ implies $\mathbf{H}^T = (\mathbf{Q}^T\mathbf{A}\mathbf{Q})^T = \mathbf{H}$, and symmetric Hessenberg matrices are tridiagonal in structure. That is,

$$\mathbf{H} = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{pmatrix} \quad \text{when } \mathbf{A} = \mathbf{A}^T. \tag{7.11.10}$$

This makes $\mathbf{Q}$ particularly easy to determine. While the matrix $\mathbf{Q}$ in (7.11.9) was only $n \times k$, let's be greedy and look for an $n \times n$ orthogonal matrix $\mathbf{Q}$ such that $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{H}$, where $\mathbf{H}$ is tridiagonal as depicted in (7.11.10). If we set $\mathbf{Q} = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\big)$, and if we agree to let $\beta_0 = 0$ and $\mathbf{q}_{n+1} = \mathbf{0}$, then

[87] Cornelius Lanczos (1893–1974) was born Kornél Löwy in Budapest, Hungary, to Jewish parents, but he changed his name to avoid trouble during the dangerous times preceding World War II. After receiving his doctorate from the University of Budapest in 1921, Lanczos moved to Germany where he became Einstein's assistant in Berlin in 1928. After coming home to Germany from a visit to Purdue University in Lafayette, Indiana, in 1931, Lanczos decided that the political climate in Germany was unacceptable, and he returned to Purdue in 1932 to continue his work in mathematical physics. The development of electronic computers stimulated Lanczos's interest in numerical analysis, and this led to positions at the Boeing Company in Seattle and at the Institute for Numerical Analysis of the National Bureau of Standards in Los Angeles. When senator Joseph R. McCarthy led a crusade against communism in the 1950s, Lanczos again felt threatened, so he left the United States to accept an offer from the famous Nobel physicist Erwin Schrödinger (1887–1961) to head the Theoretical Physics Department at the Dublin Institute for Advanced Study in Ireland where Lanczos returned to his first love—the theory of relativity. Lanczos was aware of the fast Fourier transform algorithm (p. 373) 25 years before the heralded work of J. W. Cooley and J. W. Tukey (p. 368) in 1965, but 1940 was too early for applications of the FFT to be realized. This is yet another instance where credit and fame are accorded to those who first make good use of an idea rather than to those who first conceive it.

equating the jth column of $\mathbf{AQ}$ to the jth column of $\mathbf{QH}$ tells us that we must have

$$\mathbf{A}\mathbf{q}_j = \beta_{j-1}\mathbf{q}_{j-1} + \alpha_j\mathbf{q}_j + \beta_j\mathbf{q}_{j+1} \quad \text{for } j = 1, 2, \ldots, n$$

or, equivalently,

$$\beta_j\mathbf{q}_{j+1} = \mathbf{v}_j, \quad \text{where} \quad \mathbf{v}_j = \mathbf{A}\mathbf{q}_j - \alpha_j\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1} \quad \text{for } j = 1, 2, \ldots, n.$$

By observing that $\alpha_j = \mathbf{q}_j^T\mathbf{A}\mathbf{q}_j$ and $\beta_j = \|\mathbf{v}_j\|_2$, we are led to ***Lanczos's algorithm.***

- Start with an arbitrary $\mathbf{b} \neq \mathbf{0}$, set $\beta_0 = 0$, $\mathbf{q}_0 = \mathbf{0}$, $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and iterate as indicated below.

$$\begin{array}{l}
\textit{For } j = 1 \textit{ to } n \\
\quad \mathbf{v} \leftarrow \mathbf{A}\mathbf{q}_j \\
\quad \alpha_j \leftarrow \mathbf{q}_j^T\mathbf{v} \\
\quad \mathbf{v} \leftarrow \mathbf{v} - \alpha_j\mathbf{q}_j - \beta_{j-1}\mathbf{q}_{j-1} \\
\quad \beta_j \leftarrow \|\mathbf{v}\|_2 \\
\quad \textit{If } \beta_j = 0, \textit{ then quit} \\
\quad \mathbf{q}_{j+1} \leftarrow \mathbf{v}/\beta_j \\
\textit{End}
\end{array}$$

After the kth step we have an $n \times (k+1)$ matrix $\mathbf{Q}_{k+1} = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_{k+1}\big)$ of orthonormal columns such that

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1}\begin{pmatrix} \mathbf{T}_k \\ \beta_k\mathbf{e}_k^T \end{pmatrix}, \text{ where } \mathbf{T}_k \text{ is the } k \times k \text{ tridiagonal form (7.11.10).}$$

If the iteration terminates prematurely because $\beta_j = 0$ for $j < n$, then restart the algorithm with a new initial vector $\mathbf{b}$ that is orthogonal to $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_j$. When a full orthonormal set $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ has been computed and turned into an orthogonal matrix $\mathbf{Q}$, we will have

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{T}_m \end{pmatrix}, \quad \text{where each } \mathbf{T}_i \text{ is tridiagonal} \tag{7.11.11}$$

with the splits occurring at rows where the β_j 's are zero. Of course, having these splits is generally a desirable state of affairs, especially when the objective is to compute the eigenvalues of $\mathbf{A}$.

Note: The Lanczos algorithm is computationally efficient because if each row of $\mathbf{A}$ has ν nonzero entries, then each matrix–vector product uses νn multiplications, so each step of the process uses only $\nu n + 4n$ multiplications (and about

the same number of additions). This can be a tremendous savings over what is required by Householder (or Givens) reduction as discussed in Example 5.7.4 (p. 350). Once the form (7.11.11) has been determined, spectral properties of $\mathbf{A}$ usually can be extracted by a variety of standard methods such as the QR iteration (p. 535). An alternative to computing the full tridiagonal decomposition is to stop the Lanczos iteration before completion, accept the Ritz values (the eigenvalues $\mathbf{H}_{k\times k} = \mathbf{Q}_{k\times n}^T\mathbf{A}\mathbf{Q}_{n\times k}$) as approximations to a portion of $\sigma(\mathbf{A})$, deflate the problem, and repeat the process on the smaller result.

Even when $\mathbf{A}$ is not symmetric, the same logic that produces the Lanczos algorithm can be applied to obtain an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{H}$ is upper Hessenberg. But we can't expect to obtain the efficiency that Lanczos provides because the tridiagonal structure is lost. The more general algorithm is called ***Arnoldi's*** [88] ***method,*** and it's presented below.

Example 7.11.5

Arnoldi Orthogonalization Algorithm. Given $\mathbf{A} \in \mathcal{C}^{n\times n}$, the goal is to compute an orthogonal matrix $\mathbf{Q} = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_n\big)$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{H}$ is upper Hessenberg. Proceed in the manner that produced the Lanczos algorithm by equating the jth column of $\mathbf{AQ}$ to the jth column of $\mathbf{QH}$ to obtain

$$\mathbf{A}\mathbf{q}_j = \sum_{i=1}^{j+1}\mathbf{q}_i h_{ij} \quad\Longrightarrow\quad \mathbf{q}_k^T\mathbf{A}\mathbf{q}_j = \sum_{i=1}^{j+1}\mathbf{q}_k^T\mathbf{q}_i h_{ij} = h_{kj} \quad \text{for each } 1 \le k \le j$$

$$\Longrightarrow\quad h_{j+1,j}\mathbf{q}_{j+1} = \mathbf{A}\mathbf{q}_j - \sum_{i=1}^{j}\mathbf{q}_i h_{ij}.$$

By observing that $h_{j+1,j} = \|\mathbf{v}_j\|_2$ for $\mathbf{v}_j = \mathbf{A}\mathbf{q}_j - \sum_{i=1}^{j}\mathbf{q}_i h_{ij}$, we are led to ***Arnoldi's algorithm.***

- Start with an arbitrary $\mathbf{b} \neq \mathbf{0}$, set $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and then iterate as indicated below.

---

[88] Walter Edwin Arnoldi (1917–1995) was an American engineer who published this technique in 1951, not far from the time that Lanczos's algorithm emerged. Arnoldi received his undergraduate degree in mechanical engineering from Stevens Institute of Technology, Hoboken, New Jersey, in 1937 and his MS degree at Harvard University in 1939. He spent his career working as an engineer in the Hamilton Standard Division of the United Aircraft Corporation where he eventually became the division's chief researcher. He retired in 1977. While his research concerned mechanical and aerodynamic properties of aircraft and aerospace structures, Arnoldi's name is kept alive by his orthogonalization procedure.

$$
\begin{array}{l}
\textit{For } j = 1 \textit{ to } n \\
\quad \mathbf{v} \quad \leftarrow \quad \mathbf{A}\mathbf{q}_j \\
\quad \textit{For } i = 1 \textit{ to } j \\
\qquad h_{ij} \leftarrow \mathbf{q}_i^T \mathbf{v} \\
\qquad \mathbf{v} \leftarrow \mathbf{v} - h_{ij}\mathbf{q}_i \\
\quad \textit{End For} \\
\quad h_{j+1,j} \leftarrow \|\mathbf{v}\|_2 \\
\quad \textit{If } h_{j+1,j} = 0, \textit{ then quit} \\
\quad \mathbf{q}_{j+1} \leftarrow \mathbf{v}/h_{j+1,j} \\
\textit{End For}
\end{array}
\tag{7.11.12}
$$

After the kth step we have an $n \times (k+1)$ matrix $\mathbf{Q}_{k+1} = \big(\mathbf{q}_1 \,|\, \mathbf{q}_2 \,|\, \cdots \,|\, \mathbf{q}_{k+1}\big)$ of orthonormal columns such that

$$
\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_{k+1} \begin{pmatrix} \mathbf{H}_k \\ h_{k+1,k}\mathbf{e}_k^T \end{pmatrix}, \tag{7.11.13}
$$

where $\mathbf{H}_k$ is a $k \times k$ upper-Hessenberg matrix.

Note: Remarks similar to those made about the Lanczos algorithm also hold for Arnoldi's algorithm, but the computational efficiency of Arnoldi is not as great as that of Lanczos. Close examination of Arnoldi's method reveals that it amounts to a modified Gram–Schmidt process (p. 316).

Krylov methods are a natural way to solve systems of linear equations. To see why, suppose that $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is a nonsingular system, and let $v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ be the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$. Since $\alpha_0 \neq 0$ (otherwise $v(x)/x$ would be an annihilating polynomial for $\mathbf{b}$ of degree less than $\deg v$), we have

$$
\mathbf{A}^k\mathbf{b} - \sum_{j=0}^{k-1} \alpha_j \mathbf{A}^j \mathbf{b} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{A}\left[\frac{\mathbf{A}^{k-1}\mathbf{b} - \alpha_{k-1}\mathbf{A}^{k-2}\mathbf{b} - \cdots - \alpha_1\mathbf{b}}{\alpha_0}\right] = \mathbf{b}.
$$

In other words, the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is somewhere in the Krylov space $\mathcal{K}_k$.

A technique for sorting through $\mathcal{K}_k$ to find the solution (or at least an acceptable approximate solution) of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is to sequentially consider the subspaces $\mathbf{A}(\mathcal{K}_1), \mathbf{A}(\mathcal{K}_2), \ldots, \mathbf{A}(\mathcal{K}_k)$, where at the jth step of the process the vector $\mathbf{x}_j \in \mathbf{A}(\mathcal{K}_j)$ that is closest to $\mathbf{b}$ is used as an approximation to $\mathbf{x}$. If $\mathbf{Q}_j$ is an $n \times j$ orthogonal matrix whose columns constitute a basis for $\mathcal{K}_j$, then $R(\mathbf{A}\mathbf{Q}_j) = \mathbf{A}(\mathcal{K}_j)$, so the vector $\mathbf{x}_j \in \mathbf{A}(\mathcal{K}_j)$ that is closest to $\mathbf{b}$ is the orthogonal projection of $\mathbf{b}$ onto $R(\mathbf{A}\mathbf{Q}_j)$. This means that $\mathbf{x}_j$ is the least squares solution of $\mathbf{A}\mathbf{Q}_j\mathbf{z} = \mathbf{b}$ (p. 439). If the solution of this least squares problem yields a vector $\mathbf{x}_j$ such that the residual $\mathbf{r}_j = \mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{x}_j$ is zero (or satisfactorily small), then set $\mathbf{x} = \mathbf{Q}_j\mathbf{x}_j$, and quit. Otherwise move up one

dimension, and compute the least squares solution $\mathbf{x}_{j+1}$ of $\mathbf{AQ}_{j+1}\mathbf{z} = \mathbf{b}$. Since $\mathbf{x} \in \mathcal{K}_k$, the process is guaranteed to terminate in $k \leq n$ steps or less (when exact arithmetic is used). When Arnoldi's method is used to implement this idea, the resulting algorithm is known as ***GMRES*** (an acronym for the *generalized minimal residual* algorithm that was formulated by Yousef Saad and Martin H. Schultz in 1986).

Example 7.11.6

GMRES Algorithm. To implement the idea discussed above by employing Arnoldi's algorithm, recall from (7.11.13) that after j steps of the Arnoldi process we have matrices $\mathbf{Q}_j$ and $\mathbf{Q}_{j+1}$ with orthonormal columns that span $\mathcal{K}_j$ and $\mathcal{K}_{j+1}$, respectively, along with a $j \times j$ upper-Hessenberg matrix $\mathbf{H}_j$ such that

$$\mathbf{AQ}_j = \mathbf{Q}_{j+1}\widetilde{\mathbf{H}}_j, \quad \text{where} \quad \widetilde{\mathbf{H}}_j = \begin{pmatrix} \mathbf{H}_j \\ h_{j+1,j}\mathbf{e}_j^T \end{pmatrix}.$$

Consequently the least squares solution of $\mathbf{AQ}_j\mathbf{z} = \mathbf{b}$ is the same as the least squares solution of $\mathbf{Q}_{j+1}\widetilde{\mathbf{H}}_j\mathbf{z} = \mathbf{b}$, which in turn is the same as the least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \mathbf{Q}_{j+1}^T\mathbf{b}$. But $\mathbf{Q}_{j+1}^T\mathbf{b} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ (because the first column in $\mathbf{Q}_{j+1}$ is $\mathbf{b}/\|\mathbf{b}\|_2$), so the GMRES algorithm is as follows.

- To compute the solution to a nonsingular linear system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b} \neq \mathbf{0}$, start with $\mathbf{q}_1 = \mathbf{b}/\|\mathbf{b}\|_2$, and iterate as indicated below.

  For $j = 1$ *to* n
  execute the jth Arnoldi step in (7.11.12)
  compute the least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ by using a QR factorization of $\widetilde{\mathbf{H}}_j$ (see **Note** at the end of the example)
  If $\|\mathbf{b} - \mathbf{AQ}_j\mathbf{z}\|_2 = 0$ (or is satisfactorily small)
  set $\mathbf{x} = \mathbf{Q}_j\mathbf{z}$, and quit (see **Note** at the end of the example)
  End If
  End For

The structure of the $\widetilde{\mathbf{H}}_j$'s allows us to update the QR factors of $\widetilde{\mathbf{H}}_j$ to produce the QR factors of $\widetilde{\mathbf{H}}_{j+1}$ with a single plane rotation (p. 333). To see how this is done, consider what happens when moving from the third step to the fourth step of the process. Let $\mathbf{U}_3 = \begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}$ be the 4×4 orthogonal matrix that was previously accumulated (as a product of plane rotations) to give $\mathbf{U}_3\widetilde{\mathbf{H}}_3 = \begin{pmatrix} \mathbf{R}_3 \\ \mathbf{0} \end{pmatrix}$ with $\mathbf{R}_3$ being upper triangular so that $\widetilde{\mathbf{H}}_3 = \mathbf{QR}_3$. Since

$$\begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \widetilde{\mathbf{H}}_4 = \begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} \left(\begin{array}{ccc|c} & \widetilde{\mathbf{H}}_3 & & \begin{matrix}\star\\ \star\\ \star\\ \star\end{matrix} \\ \hline 0 & 0 & 0 & \star \end{array}\right) = \left(\begin{array}{ccc|c} & \mathbf{U}_3\widetilde{\mathbf{H}}_3 & & \begin{matrix}\star\\ \star\\ \star\\ \star\end{matrix} \\ \hline 0 & 0 & 0 & \star \end{array}\right) = \left(\begin{array}{cccc} \star & \star & \star & \star \\ 0 & \star & \star & \star \\ 0 & 0 & \star & \star \\ 0 & 0 & 0 & \star \\ \hline 0 & 0 & 0 & \star \end{array}\right),$$

a plane rotation of the form $\mathbf{P}_{45} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & c & s \\ & & & -s & c \end{pmatrix}$ will annihilate the entry in the lower-right-hand corner of this last array. Consequently, $\mathbf{U}_4 = \mathbf{P}_{45}\begin{pmatrix} \mathbf{U}_3 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}$ is an orthogonal matrix such that $\mathbf{U}_4\widetilde{\mathbf{H}}_4 = \begin{pmatrix} \mathbf{R}_4 \\ \mathbf{0} \end{pmatrix}$, where $\mathbf{R}_4$ is upper triangular, and this produces the QR factors of $\widetilde{\mathbf{H}}_4$.

Note: The value of the residual norm $\|\mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{z}\|_2$ at each step of GMRES is available at almost no cost. To see why, notice that the previous discussion shows that at the jth step there is a $(j+1) \times (j+1)$ orthogonal matrix $\mathbf{U} = \begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}$ (that exists as an accumulation of plane rotations) such that $\mathbf{U}\widetilde{\mathbf{H}}_j = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}$, and this produces $\widetilde{\mathbf{H}}_j = \mathbf{QR}$. The least squares solution of $\widetilde{\mathbf{H}}_j\mathbf{z} = \|\mathbf{b}\|_2\,\mathbf{e}_1$ is obtained by solving $\mathbf{Rz} = \mathbf{Q}^T\,\|\mathbf{b}\|_2\,\mathbf{e}_1$ (p. 314), so

$$\begin{aligned} \|\mathbf{b} - \mathbf{A}\mathbf{Q}_j\mathbf{z}\|_2 &= \left\| \|\mathbf{b}\|_2\,\mathbf{e}_1 - \widetilde{\mathbf{H}}_j\mathbf{z} \right\|_2 = \left\| \|\mathbf{b}\|_2\,\mathbf{U}\mathbf{e}_1 - \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}\mathbf{z} \right\|_2 \\ &= \left\| \|\mathbf{b}\|_2 \begin{pmatrix} \mathbf{Q}^T \\ \mathbf{v}^T \end{pmatrix}\mathbf{e}_1 - \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}\mathbf{z} \right\|_2 = \left\| \begin{pmatrix} \mathbf{0} \\ \|\mathbf{b}\|_2\,\mathbf{v}^T\mathbf{e}_1 \end{pmatrix} \right\|_2 \\ &= \|\mathbf{b}\|_2\,|u_{j+1,1}|. \end{aligned}$$

Since $u_{j+1,1}$ is just the last entry in the accumulation of the various plane rotations applied to $\mathbf{e}_1$, the cost of producing these values as the algorithm proceeds is small, so deciding on the acceptability of an approximate solution at each step in the GMRES algorithm is cheap.

When solving nonsingular symmetric systems $\mathbf{Ax} = \mathbf{b}$, a strategy similar to the one that produced the GMRES algorithm can be adopted except that the Lanczos procedure (p. 651) is used in place of the Arnoldi process (p. 653). When this is done, the resulting algorithm is called ***MINRES*** (an acronym for *minimal residual algorithm*), and, as you might guess, there is an increase in computational efficiency when Lanczos replaces Arnoldi. Historically, MINRES preceded GMRES.

Another Krylov method that deserves mention is the ***conjugate gradient algorithm***, presented by Magnus R. Hestenes and Eduard Stiefel in 1952, that is used to solve positive definite systems.

Example 7.11.7

Conjugate Gradient Algorithm. Suppose that $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b} \neq \mathbf{0}$ is a (real) positive definite system, and suppose that the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$ is $v(x) = x^k - \sum_{j=0}^{k-1} \alpha_j x^j$ so that the solution $\mathbf{x}$ is somewhere in the Krylov space $\mathcal{K}_k$ (p. 654). The conjugate gradient algorithm emanated from the observation that if $\mathbf{A}$ is positive definite, then the quadratic function

$$f(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{2} - \mathbf{x}^T\mathbf{b}$$

has as its gradient

$$\nabla f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b},$$

and there is a unique minimizer for f that happens to be the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Consequently, any technique that attempts to minimize f is a technique that attempts to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$. Since the $\mathbf{x}$ is somewhere in $\mathcal{K}_k$, it makes sense to try to minimize f over $\mathcal{K}_k$. One approach for doing this is the ***method of steepest descent*** in which a current approximation $\mathbf{x}_j$ is updated by adding a correction term directed along the negative gradient $-\nabla f(\mathbf{x}_j) = \mathbf{b} - \mathbf{A}\mathbf{x}_j = \mathbf{r}_j$ (the jth residual). In other words, let

$$\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \mathbf{r}_j, \quad \text{and set} \quad \alpha_j = \frac{\mathbf{r}_j^T\mathbf{r}_j}{\mathbf{r}_j^T\mathbf{A}\mathbf{r}_j}$$

because this α_j minimizes $f(\mathbf{x}_{j+1})$. In spite of the fact that successive residuals are orthogonal ($\mathbf{r}_{j+1}^T\mathbf{r}_j = \mathbf{0}$), the rate of convergence can be slow because as the ratio of eigenvalues $\lambda_{\max}(\mathbf{A})/\lambda_{\min}(\mathbf{A})$ becomes larger, the surface defined by f becomes more distorted, and a negative gradient $\mathbf{r}_j$ need not point in a direction aimed anywhere near the lowest point on the surface. An ingenious mechanism for overcoming this difficulty is to replace the search directions $\mathbf{r}_j$ by directions defined by vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots$ that are ***conjugate*** to each other in the sense that $\mathbf{q}_i^T\mathbf{A}\mathbf{q}_j = 0$ for all $i \neq j$ (some authors say "A-orthogonal"). Starting with $\mathbf{x}_0 = \mathbf{0}$, the idea is to begin by moving in the direction of steepest descent with

$$\mathbf{x}_1 = \alpha_1\mathbf{q}_1, \quad \text{where } \mathbf{q}_1 = \mathbf{r}_0 = \mathbf{b} \text{ and } \alpha_1 = \frac{\mathbf{r}_0^T\mathbf{r}_0}{\mathbf{r}_0^T\mathbf{A}\mathbf{r}_0},$$

but at the second step use a direction vector

$$\mathbf{q}_2 = \mathbf{r}_1 + \beta_1\mathbf{q}_1, \quad \text{where } \beta_1 \text{ is chosen to force } \mathbf{q}_2^T\mathbf{A}\mathbf{q}_1 = 0.$$

With a bit of effort you can see that $\beta_1 = \mathbf{r}_1^T\mathbf{r}_1/\mathbf{r}_0^T\mathbf{r}_0$ does the job. Then set $\mathbf{x}_2 = \mathbf{x}_1 + \alpha_2\mathbf{q}_2$, and recycle the process. The formal algorithm is as follows.

Formal Conjugate Gradient Algorithm. To compute the solution to a positive definite linear system $\mathbf{A}_{n\times n}\mathbf{x} = \mathbf{b}$, start with $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{b}$, and $\mathbf{q}_1 = \mathbf{b}$, and iterate as indicated below.

For $j = 1$ *to* n

$\alpha_j \leftarrow \mathbf{r}_{j-1}^T\mathbf{r}_{j-1}/\mathbf{q}_j^T\mathbf{A}\mathbf{q}_j$ (step size)

$\mathbf{x}_j \leftarrow \mathbf{x}_{j-1} + \alpha_j\mathbf{q}_j$ (approximate solution)

$\mathbf{r}_j \leftarrow \mathbf{r}_{j-1} - \alpha_j\mathbf{A}\mathbf{q}_j$ (residual)

If $\|\mathbf{r}_j\|_2 = 0$ (or is satisfactorily small)

set $\mathbf{x} = \mathbf{x}_j$, and quit

End If

$\beta_j \leftarrow \mathbf{r}_j^T\mathbf{r}_j/\mathbf{r}_{j-1}^T\mathbf{r}_{j-1}$ (conjugation factor)

$\mathbf{q}_{j+1} \leftarrow \mathbf{r}_j + \beta_j\mathbf{q}_j$ (search direction)

End For

It can be shown that vectors produced by this algorithm after j steps are such that (in exact arithmetic)

$$span\,\{\mathbf{x}_1, \ldots, \mathbf{x}_j\} = span\,\{\mathbf{q}_1, \ldots, \mathbf{q}_j\} = span\,\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{j-1}\} = \mathcal{K}_j,$$

and, in addition to having $\mathbf{q}_i\mathbf{A}\mathbf{q}_j = 0$ for $i < j$, the residuals are orthogonal—i.e., $\mathbf{r}_i^T\mathbf{r}_j = 0$ for $i < j$. Furthermore, the algorithm will find the solution in $k \le n$ steps.

As mentioned earlier, Krylov solvers such as GMRES and the conjugate gradient algorithm produce the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ in $k \le n$ steps (in exact arithmetic), so, at first glance, this looks like good news. But in practice n can be prohibitively large, and it's not rare to have $k = n$. Consequently, Krylov algorithms are often viewed as iterative methods that are terminated long before n steps have been completed. The challenge in applying Krylov solvers (as well as iterative methods in general) revolves around the issue of how to replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ with an equivalent ***preconditioned system*** $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$ that requires only a small number of iterations to deliver a reasonably accurate approximate solution. Building effective preconditioners $\mathbf{M}^{-1}$ is part science and part art, and the techniques vary from algorithm to algorithm.

Classical linear stationary iterative methods (p. 620) are formed by splitting $\mathbf{A} = \mathbf{M} - \mathbf{N}$ and setting $\mathbf{x}(k) = \mathbf{H}\mathbf{x}(k-1) + \mathbf{d}$, where $\mathbf{H} = \mathbf{M}^{-1}\mathbf{N}$ and $\mathbf{d} = \mathbf{M}^{-1}\mathbf{b}$. This is a preconditioning technique because the effect is to replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ by $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$, where $\mathbf{M}^{-1}\mathbf{A} = \mathbf{I} - \mathbf{H}$ such that $\rho(\mathbf{H}) < 1$. The goal is to find an easily inverted $\mathbf{M}$ (in the sense that $\mathbf{M}\mathbf{d} = \mathbf{b}$ is easily solved) that drives the value of $\rho(\mathbf{H})$ down far enough to insure a satisfactory rate of convergence, and this is a delicate balancing act.

The goal in preconditioning Krylov solvers is somewhat different. For example, if $k = \deg v(x)$ is the degree of the minimum polynomial of $\mathbf{b}$ with respect to $\mathbf{A}$, then GMRES sorts through $\mathcal{K}_k$ to find the solution of $\mathbf{Ax} = \mathbf{b}$ in k steps. So the aim of preconditioning GMRES might be to manipulate the interplay between $\mathbf{M}^{-1}\mathbf{b}$ and $\mathbf{M}^{-1}\mathbf{A}$ to insure that the degree of minimum polynomial $\widetilde{v}(x)$ of $\mathbf{M}^{-1}\mathbf{b}$ with respect to $\mathbf{M}^{-1}\mathbf{A}$ is significantly smaller than k. Since this is difficult to do, an alternate goal is to try to reduce the degree of the minimum polynomial $\widetilde{m}(x)$ for $\mathbf{M}^{-1}\mathbf{A}$ because driving down $\deg \widetilde{m}(x)$ also drives down $\deg \widetilde{v}(x)$—remember, $\widetilde{v}(x)$ is a divisor of $\widetilde{m}(x)$ (p. 647). If a preconditioner $\mathbf{M}^{-1}$ can be found to force $\mathbf{M}^{-1}\mathbf{A}$ to be diagonalizable with only a few distinct eigenvalues (say j of them), then $\deg \widetilde{m}(x) = j$ (p. 645), and GMRES will find the solution in no more than j steps. But this too is an overly ambitious goal for practical problems. In reality this objective is compromised by looking for a preconditioner such that $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable whose eigenvalues fall into a few small clusters—say j of them. The hope is that if $\mathbf{M}^{-1}\mathbf{A}$ is diagonalizable, and if the diameters of the clusters are small enough, then $\mathbf{M}^{-1}\mathbf{A}$ will behave numerically like a diagonalizable matrix with j distinct eigenvalues, so GMRES is inclined to produce reasonably accurate approximations in no more than j steps. While the intuition is simple, subtleties involving the magnitudes of eigenvalues, separation of clusters, and the meaning of "small diameter" complicate the picture to make definitive statements and rigorous arguments difficult to formulate. Constructing good preconditioners and proving they actually work as advertised remains an active area of research in the field of numerical analysis.

Only the tip of the iceberg concerning practical applications of Krylov methods is revealed in this section. The analysis required to more fully understand the numerical behavior of various Krylov methods can be found in several excellent advanced texts specializing in matrix computations.

Exercises for section 7.11

7.11.1. Determine the minimum polynomial for $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$.

7.11.2. Find the minimum polynomial of $\mathbf{b} = (-1, 1, 1)^T$ with respect to the matrix $\mathbf{A}$ given in Exercise 7.11.1.

7.11.3. Use Krylov's method to determine the characteristic polynomial for the matrix $\mathbf{A}$ given in Exercise 7.11.1.

7.11.4. What is the Jordan form for a matrix whose minimum polynomial is $m(x) = (x - \lambda)(x - \mu)^2$ and whose characteristic polynomial is $c(x) = (x - \lambda)^2(x - \mu)^4$?

7.11.5. Use the technique described in Example 7.11.1 (p. 643) to determine the minimum polynomial for $\mathbf{A} = \begin{pmatrix} -7 & -4 & 8 & -8 \\ -4 & -1 & 4 & -4 \\ -16 & -8 & 17 & -16 \\ -6 & -3 & 6 & -5 \end{pmatrix}$.

7.11.6. Explain why similar matrices have the same minimum and characteristic polynomials.

7.11.7. Show that two matrices can have the same minimum and characteristic polynomials without being similar by considering $\mathbf{A} = \begin{pmatrix} \mathbf{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} \mathbf{N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, where $\mathbf{N} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

7.11.8. Prove that if $\mathbf{A}$ and $\mathbf{B}$ are nonderogatory matrices that have the same characteristic polynomial, then $\mathbf{A}$ is similar to $\mathbf{B}$.

7.11.9. Use the Lanczos algorithm to find an orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{T}$ is tridiagonal, where $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$.

7.11.10. Starting with $\mathbf{x}_0 = \mathbf{0}$, apply the conjugate gradient algorithm to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$.

7.11.11. Use Arnoldi's algorithm to find an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{H}$ is upper Hessenberg, where $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$.

7.11.12. Use GMRES to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{A} = \begin{pmatrix} 5 & 1 & 2 \\ -4 & 0 & -2 \\ -4 & -1 & -1 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$.

CHAPTER 8

Perron–Frobenius Theory of Nonnegative Matrices

8.1 INTRODUCTION

$\mathbf{A} \in \Re^{m\times n}$ is said to be a ***nonnegative matrix*** whenever each $a_{ij} \geq 0$, and this is denoted by writing $\mathbf{A} \geq \mathbf{0}$. In general, $\mathbf{A} \geq \mathbf{B}$ means that each $a_{ij} \geq b_{ij}$. Similarly, $\mathbf{A}$ is a ***positive matrix*** when each $a_{ij} > 0$, and this is denoted by writing $\mathbf{A} > \mathbf{0}$. More generally, $\mathbf{A} > \mathbf{B}$ means that each $a_{ij} > b_{ij}$.

Applications abound with nonnegative and positive matrices. In fact, many of the applications considered in this text involve nonnegative matrices. For example, the connectivity matrix $\mathbf{C}$ in Example 3.5.2 (p. 100) is nonnegative. The discrete Laplacian $\mathbf{L}$ from Example 7.6.2 (p. 563) leads to a nonnegative matrix because $(4\mathbf{I} - \mathbf{L}) \geq \mathbf{0}$. The matrix $e^{\mathbf{A}t}$ that defines the solution of the system of differential equations in the mixing problem of Example 7.9.7 (p. 610) is nonnegative for all $t \geq 0$. And the system of difference equations $\mathbf{p}(k) = \mathbf{A}\mathbf{p}(k-1)$ resulting from the shell game of Example 7.10.8 (p. 635) has a nonnegative coefficient matrix $\mathbf{A}$.

Since nonnegative matrices are pervasive, it's natural to investigate their properties, and that's the purpose of this chapter. A primary issue concerns the extent to which the properties $\mathbf{A} > \mathbf{0}$ or $\mathbf{A} \geq \mathbf{0}$ translate to spectral properties—e.g., to what extent does $\mathbf{A}$ have positive (or nonnegative) eigenvalues and eigenvectors?

The topic is called the "Perron–Frobenius theory" because it evolved from the contributions of the German mathematicians Oskar (or Oscar) Perron[89] and

[89] Oskar Perron (1880–1975) originally set out to fulfill his father's wishes to be in the family busi-

Ferdinand Georg Frobenius.[90] Perron published his treatment of positive matrices in 1907, and in 1912 Frobenius contributed substantial extensions of Perron's results to cover the case of nonnegative matrices.

In addition to saying something useful, the Perron–Frobenius theory is elegant. It is a testament to the fact that beautiful mathematics eventually tends to be useful, and useful mathematics eventually tends to be beautiful.

---

ness, so he only studied mathematics in his spare time. But he was eventually captured by the subject, and, after studying at Berlin, Tübingen, and Göttingen, he completed his doctorate, writing on geometry, at the University of Munich under the direction of Carl von Lindemann (1852–1939) (who first proved that π was transcendental). Upon graduation in 1906, Perron held positions at Munich, Tübingen, and Heidelberg. Perron's career was interrupted in 1915 by World War I in which he earned the Iron Cross. After the war he resumed work at Heidelberg, but in 1922 he returned to Munich to accept a chair in mathematics, a position he occupied for the rest of his career. In addition to his contributions to matrix theory, Perron's work covered a wide range of other topics in algebra, analysis, differential equations, continued fractions, geometry, and number theory. He was a man of extraordinary mental and physical energy. In addition to being able to climb mountains until he was in his midseventies, Perron continued to teach at Munich until he was 80 (although he formally retired at age 71), and he maintained a remarkably energetic research program into his nineties. He published 18 of his 218 papers *after* he was 84.

90 Ferdinand Georg Frobenius (1849–1917) earned his doctorate under the supervision of Karl Weierstrass (p. 589) at the University of Berlin in 1870. As mentioned earlier, Frobenius was a mentor to and a collaborator with Issai Schur (p. 123), and, in addition to their joint work in group theory, they were among the first to study matrix theory as a discipline unto itself. Frobenius in particular must be considered along with Cayley and Sylvester when thinking of core developers of matrix theory. However, in the beginning, Frobenius's motivation came from Kronecker (p. 597) and Weierstrass, and he seemed oblivious to Cayley's work (p. 80). It was not until 1896 that Frobenius became aware of Cayley's 1857 work, *A Memoir on the Theory of Matrices,* and only then did the terminology "matrix" appear in Frobenius's work. Even though Frobenius was the first to give a rigorous proof of the Cayley–Hamilton theorem (p. 509), he generously attributed it to Cayley in spite of the fact that Cayley had only discussed the result for 2×2 and 3×3 matrices. But credit in this regard is not overly missed because Frobenius's extension of Perron's results are more substantial, and they alone may keep Frobenius's name alive forever.

8.2 POSITIVE MATRICES

The purpose of this section is to focus on matrices $\mathbf{A}_{n\times n} > \mathbf{0}$ with positive entries, and the aim is to investigate the extent to which this positivity is inherited by the eigenvalues and eigenvectors of $\mathbf{A}$.

There are a few elementary observations that will help along the way, so let's begin with them. First, notice that

$$\mathbf{A} > \mathbf{0} \implies \rho(\mathbf{A}) > 0 \tag{8.2.1}$$

because if $\sigma(\mathbf{A}) = \{0\}$, then the Jordan form for $\mathbf{A}$, and hence $\mathbf{A}$ itself, is nilpotent, which is impossible when each $a_{ij} > 0$. This means that our discussions can be limited to positive matrices having spectral radius 1 because $\mathbf{A}$ can always be normalized by its spectral radius—i.e., $\mathbf{A} > \mathbf{0} \iff \mathbf{A}/\rho(\mathbf{A}) > \mathbf{0}$, and $\rho(\mathbf{A}) = r \iff \rho(\mathbf{A}/r) = 1$. Other easily verified observations are

$$\mathbf{P} > \mathbf{0},\ \mathbf{x} \geq \mathbf{0},\ \mathbf{x} \neq \mathbf{0} \implies \mathbf{P}\mathbf{x} > \mathbf{0}, \tag{8.2.2}$$

$$\mathbf{N} \geq \mathbf{0},\ \mathbf{u} \geq \mathbf{v} \geq \mathbf{0} \implies \mathbf{N}\mathbf{u} \geq \mathbf{N}\mathbf{v}, \tag{8.2.3}$$

$$\mathbf{N} \geq \mathbf{0},\ \mathbf{z} > \mathbf{0},\ \mathbf{N}\mathbf{z} = \mathbf{0} \implies \mathbf{N} = \mathbf{0}, \tag{8.2.4}$$

$$\mathbf{N} \geq \mathbf{0},\ \mathbf{N} \neq \mathbf{0},\ \mathbf{u} > \mathbf{v} > \mathbf{0} \implies \mathbf{N}\mathbf{u} > \mathbf{N}\mathbf{v}. \tag{8.2.5}$$

In all that follows, the bar notation $|\star|$ is used to denote a matrix of absolute values—i.e., $|\mathbf{M}|$ is the matrix having entries $|m_{ij}|$. The bar notation will *never* denote a determinant in the sequel. Finally, notice that as a simple consequence of the triangle inequality, it's always true that $|\mathbf{A}\mathbf{x}| \leq |\mathbf{A}|\,|\mathbf{x}|$.

Positive Eigenpair

If $\mathbf{A}_{n\times n} > \mathbf{0}$, then the following statements are true.

- $\rho(\mathbf{A}) \in \sigma(\mathbf{A})$. (8.2.6)
- If $\mathbf{A}\mathbf{x} = \rho(\mathbf{A})\,\mathbf{x}$, then $\mathbf{A}|\mathbf{x}| = \rho(\mathbf{A})\,|\mathbf{x}|$ and $|\mathbf{x}| > \mathbf{0}$. (8.2.7)

In other words, $\mathbf{A}$ has an eigenpair of the form $(\rho(\mathbf{A}), \mathbf{v})$ with $\mathbf{v} > 0$.

Proof. As mentioned earlier, it can be assumed that $\rho(\mathbf{A}) = 1$ without any loss of generality. If $(\lambda, \mathbf{x})$ is any eigenpair for $\mathbf{A}$ such that $|\lambda| = 1$, then

$$|\mathbf{x}| = |\lambda|\,|\mathbf{x}| = |\lambda\mathbf{x}| = |\mathbf{A}\mathbf{x}| \leq |\mathbf{A}|\,|\mathbf{x}| = \mathbf{A}\,|\mathbf{x}| \implies |\mathbf{x}| \leq \mathbf{A}\,|\mathbf{x}|. \tag{8.2.8}$$

The goal is to show that equality holds. For convenience, let $\mathbf{z} = \mathbf{A}\,|\mathbf{x}|$ and $\mathbf{y} = \mathbf{z} - |\mathbf{x}|$, and notice that (8.2.8) implies $\mathbf{y} \geq \mathbf{0}$. Suppose that $\mathbf{y} \neq \mathbf{0}$—i.e.,

suppose that some $y_i > 0$. In this case, it follows from (8.2.2) that $\mathbf{Ay} > \mathbf{0}$ and $\mathbf{z} > \mathbf{0}$, so there must exist a number $\epsilon > 0$ such that $\mathbf{Ay} > \epsilon\,\mathbf{z}$ or, equivalently,

$$\frac{\mathbf{A}}{1+\epsilon}\mathbf{z} > \mathbf{z}.$$

Writing this inequality as $\mathbf{Bz} > \mathbf{z}$, where $\mathbf{B} = \mathbf{A}/(1+\epsilon)$, and successively multiplying both sides by $\mathbf{B}$ while using (8.2.5) produces

$$\mathbf{B}^2\mathbf{z} > \mathbf{Bz} > \mathbf{z}, \quad \mathbf{B}^3\mathbf{z} > \mathbf{B}^2\mathbf{z} > \mathbf{z}, \quad \ldots \implies \mathbf{B}^k\mathbf{z} > \mathbf{z} \quad \text{for all } k = 1, 2, \ldots.$$

But $\lim_{k\to\infty} \mathbf{B}^k = \mathbf{0}$ because $\rho(\mathbf{B}) = \sigma\big(\mathbf{A}/(1+\epsilon)\big) = 1/(1+\epsilon) < 1$ (recall (7.10.5) on p. 617), so, in the limit, we have $\mathbf{0} > \mathbf{z}$, which contradicts the fact that $\mathbf{z} > \mathbf{0}$. Since the supposition that $\mathbf{y} \neq \mathbf{0}$ led to this contradiction, the supposition must be false and, consequently, $\mathbf{0} = \mathbf{y} = \mathbf{A}\,|\mathbf{x}| - |\mathbf{x}|$. Thus $|\mathbf{x}|$ is an eigenvector for $\mathbf{A}$ associated with the eigenvalue $1 = \rho(\mathbf{A})$. The proof is completed by observing that $|\mathbf{x}| = \mathbf{A}\,|\mathbf{x}| = \mathbf{z} > \mathbf{0}$. ∎

Now that it's been established that $\rho(\mathbf{A}) > 0$ is in fact an eigenvalue for $\mathbf{A} > \mathbf{0}$, the next step is to investigate the index of this special eigenvalue.

Index of $\rho(\mathbf{A})$

If $\mathbf{A}_{n\times n} > \mathbf{0}$, then the following statements are true.

- $\rho(\mathbf{A})$ is the only eigenvalue of $\mathbf{A}$ on the spectral circle.
- $\mathit{index}(\rho(\mathbf{A})) = 1$. In other words, $\rho(\mathbf{A})$ is a *semisimple* eigenvalue. Recall Exercise 7.8.4 (p. 596).

Proof. Again, assume without loss of generality that $\rho(\mathbf{A}) = 1$. We know from (8.2.7) on p. 663 that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$ such that $|\lambda| = 1$, then $\mathbf{0} < |\mathbf{x}| = \mathbf{A}\,|\mathbf{x}|$, so $0 < |x_k| = \big(\mathbf{A}\,|\mathbf{x}|\big)_k = \sum_{j=1}^n a_{kj}|x_j|$. But it's also true that $|x_k| = |\lambda|\,|x_k| = |(\lambda\mathbf{x})_k| = |(\mathbf{Ax})_k| = \big|\sum_{j=1}^n a_{kj}x_j\big|$, and thus

$$\Big|\sum_j a_{kj}x_j\Big| = \sum_j a_{kj}|x_j| = \sum_j |a_{kj}x_j|. \tag{8.2.9}$$

For nonzero vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \subset \mathcal{C}^n$, it's a fact that $\|\sum_j \mathbf{z}_j\|_2 = \sum_j \|\mathbf{z}_j\|_2$ (equality in the triangle inequality) if and only if each $\mathbf{z}_j = \alpha_j\mathbf{z}_1$ for some $\alpha_j > 0$ (Exercise 5.1.10, p. 277). In particular, this holds for scalars, so (8.2.9) insures the existence of numbers $\alpha_j > 0$ such that

$$a_{kj}x_j = \alpha_j(a_{k1}x_1) \quad \text{or, equivalently,} \quad x_j = \pi_j x_1 \quad \text{with } \pi_j = \frac{\alpha_j a_{k1}}{a_{kj}} > 0.$$

In other words, if $|\lambda| = 1$, then $\mathbf{x} = x_1\mathbf{p}$, where $\mathbf{p} = (1, \pi_2, \ldots, \pi_n)^T > \mathbf{0}$, so

$$\lambda\mathbf{x} = \mathbf{A}\mathbf{x} \implies \lambda\mathbf{p} = \mathbf{A}\mathbf{p} = |\mathbf{A}\mathbf{p}| = |\lambda\mathbf{p}| = |\lambda|\mathbf{p} = \mathbf{p} \implies \lambda = 1,$$

and thus 1 is the only eigenvalue of $\mathbf{A}$ on the spectral circle. Now suppose that $\mathit{index}(1) = m > 1$. It follows that $\left\|\mathbf{A}^k\right\|_\infty \to \infty$ as $k \to \infty$ because there is an $m \times m$ Jordan block $\mathbf{J}_\star$ in the Jordan form $\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ that looks like (7.10.30) on p. 629, so $\left\|\mathbf{J}_\star^k\right\|_\infty \to \infty$, which in turn means that $\left\|\mathbf{J}^k\right\|_\infty \to \infty$ and, consequently, $\left\|\mathbf{J}^k\right\|_\infty = \left\|\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}\right\|_\infty \le \left\|\mathbf{P}^{-1}\right\|_\infty \left\|\mathbf{A}^k\right\|_\infty \left\|\mathbf{P}\right\|_\infty$ implies

$$\left\|\mathbf{A}^k\right\|_\infty \ge \frac{\left\|\mathbf{J}^k\right\|_\infty}{\left\|\mathbf{P}^{-1}\right\|_\infty \left\|\mathbf{P}\right\|_\infty} \to \infty.$$

Let $\mathbf{A}^k = \left[a_{ij}^{(k)}\right]$, and let i_k denote the row index for which $\left\|\mathbf{A}^k\right\|_\infty = \sum_j a_{i_k j}^{(k)}$. We know that there exists a vector $\mathbf{p} > \mathbf{0}$ such that $\mathbf{p} = \mathbf{A}\mathbf{p}$, so for such an eigenvector,

$$\|\mathbf{p}\|_\infty \ge p_{i_k} = \sum_j a_{i_k j}^{(k)} p_j \ge \left(\sum_j a_{i_k j}^{(k)}\right)(\min_i p_i) = \left\|\mathbf{A}^k\right\|_\infty (\min_i p_i) \to \infty.$$

But this is impossible because $\mathbf{p}$ is a constant vector, so the supposition that $\mathit{index}(1) > 1$ must be false, and thus $\mathit{index}(1) = 1$. ∎

Establishing that $\rho(\mathbf{A})$ is a semisimple eigenvalue of $\mathbf{A} > \mathbf{0}$ was just a steppingstone (but an important one) to get to the following theorem concerning the multiplicities of $\rho(\mathbf{A})$.

Multiplicities of $\rho(\mathbf{A})$

If $\mathbf{A}_{n\times n} > \mathbf{0}$, then $\mathit{alg\ mult}_{\mathbf{A}}(\rho(\mathbf{A})) = 1$. In other words, the spectral radius of $\mathbf{A}$ is a *simple* eigenvalue of $\mathbf{A}$.

So $\dim N(\mathbf{A} - \rho(\mathbf{A})\mathbf{I}) = \mathit{geo\ mult}_{\mathbf{A}}(\rho(\mathbf{A})) = \mathit{alg\ mult}_{\mathbf{A}}(\rho(\mathbf{A})) = 1$.

Proof. As before, assume without loss of generality that $\rho(\mathbf{A}) = 1$, and suppose that $\mathit{alg\ mult}_{\mathbf{A}}(\lambda = 1) = m > 1$. We already know that $\lambda = 1$ is a semisimple eigenvalue, which means that $\mathit{alg\ mult}_{\mathbf{A}}(1) = \mathit{geo\ mult}_{\mathbf{A}}(1)$ (p. 510), so there are m linearly independent eigenvectors associated with $\lambda = 1$. If $\mathbf{x}$ and $\mathbf{y}$ are a pair of independent eigenvectors associated with $\lambda = 1$, then $\mathbf{x} \ne \alpha\mathbf{y}$ for all $\alpha \in \mathcal{C}$. Select a nonzero component from $\mathbf{y}$, say $y_i \ne 0$, and set $\mathbf{z} = \mathbf{x} - (x_i/y_i)\mathbf{y}$. Since $\mathbf{A}\mathbf{z} = \mathbf{z}$, we know from (8.2.7) on p. 663 that $\mathbf{A}|\mathbf{z}| = |\mathbf{z}| > \mathbf{0}$. But this contradicts the fact that $z_i = x_i - (x_i/y_i)y_i = 0$. Therefore, the supposition that $m > 1$ must be false, and thus $m = 1$. ∎

Since $N(\mathbf{A} - \rho(\mathbf{A})\mathbf{I})$ is a one-dimensional space that can be spanned by some $\mathbf{v} > 0$, there is a *unique* eigenvector $\mathbf{p} \in N(\mathbf{A} - \rho(\mathbf{A})\mathbf{I})$ such that $\mathbf{p} > \mathbf{0}$ and $\sum_j p_j = 1$ (it's obtained by the normalization $\mathbf{p} = \mathbf{v}/\|\mathbf{v}\|_1$—see Exercise 8.2.3). This special eigenvector $\mathbf{p}$ is called the ***Perron vector*** for $\mathbf{A} > \mathbf{0}$, and the associated eigenvalue $r = \rho(\mathbf{A})$ is called the ***Perron root*** of $\mathbf{A}$.

Since $\mathbf{A} > \mathbf{0} \iff \mathbf{A}^T > \mathbf{0}$, and since $\rho(\mathbf{A}) = \rho(\mathbf{A}^T)$, it's clear that if $\mathbf{A} > \mathbf{0}$, then in addition to the Perron eigenpair $(r, \mathbf{p})$ for $\mathbf{A}$ there is a corresponding Perron eigenpair $(r, \mathbf{q})$ for $\mathbf{A}^T$. Because $\mathbf{q}^T\mathbf{A} = r\mathbf{q}^T$, the vector $\mathbf{q}^T > \mathbf{0}$ is called the ***left-hand Perron vector*** for $\mathbf{A}$.

While eigenvalues of $\mathbf{A} > \mathbf{0}$ other than $\rho(\mathbf{A})$ may or may not be positive, it turns out that no eigenvectors other than positive multiples of the Perron vector can be positive—or even nonnegative.

No Other Positive Eigenvectors

There are no nonnegative eigenvectors for $\mathbf{A}_{n\times n} > \mathbf{0}$ other than the Perron vector $\mathbf{p}$ and its positive multiples. (8.2.10)

Proof. If $(\lambda, \mathbf{y})$ is an eigenpair for $\mathbf{A}$ such that $\mathbf{y} \geq \mathbf{0}$, and if $\mathbf{x} > \mathbf{0}$ is the Perron vector for $\mathbf{A}^T$, then $\mathbf{x}^T\mathbf{y} > 0$ by (8.2.2), so

$$\rho(\mathbf{A})\mathbf{x}^T = \mathbf{x}^T\mathbf{A} \implies \rho(\mathbf{A})\mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{A}\mathbf{y} = \lambda\mathbf{x}^T\mathbf{y} \implies \rho(\mathbf{A}) = \lambda. \quad \blacksquare$$

In 1942 the German mathematician Lothar Collatz (1910–1990) discovered the following formula for the Perron root, and in 1950 Helmut Wielandt (p. 534) used it to develop the Perron–Frobenius theory.

Collatz–Wielandt Formula

The Perron root of $\mathbf{A}_{n\times n} > \mathbf{0}$ is given by $r = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x})$, where

$$f(\mathbf{x}) = \min_{\substack{1\leq i\leq n\\ x_i\neq 0}} \frac{[\mathbf{A}\mathbf{x}]_i}{x_i} \quad \text{and} \quad \mathcal{N} = \{\mathbf{x} \,|\, \mathbf{x} \geq \mathbf{0} \text{ with } \mathbf{x} \neq \mathbf{0}\}.$$

Proof. If $\xi = f(\mathbf{x})$ for $\mathbf{x} \in \mathcal{N}$, then $\mathbf{0} \leq \xi\mathbf{x} \leq \mathbf{A}\mathbf{x}$. Let $\mathbf{p}$ and $\mathbf{q}^T$ be the respective the right-hand and left-hand Perron vectors for $\mathbf{A}$ associated with the Perron root r, and use (8.2.3) along with $\mathbf{q}^T\mathbf{x} > 0$ (by (8.2.2)) to write

$$\xi\mathbf{x} \leq \mathbf{A}\mathbf{x} \implies \xi\mathbf{q}^T\mathbf{x} \leq \mathbf{q}^T\mathbf{A}\mathbf{x} = r\mathbf{q}^T\mathbf{x} \implies \xi \leq r \implies f(\mathbf{x}) \leq r \;\; \forall\, \mathbf{x} \in \mathcal{N}.$$

Since $f(\mathbf{p}) = r$ and $\mathbf{p} \in \mathcal{N}$, it follows that $r = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x})$. $\blacksquare$

Below is a summary of the results obtained in this section.

Perron's Theorem

If $\mathbf{A}_{n\times n} > \mathbf{0}$ with $r = \rho(\mathbf{A})$, then the following statements are true.

- $r > 0.$ (8.2.11)
- $r \in \sigma(\mathbf{A})$ (r is called the ***Perron root***). (8.2.12)
- $alg\ mult_{\mathbf{A}}(r) = 1.$ (8.2.13)
- There exists an eigenvector $\mathbf{x} > \mathbf{0}$ such that $\mathbf{A}\mathbf{x} = r\mathbf{x}.$ (8.2.14)
- The ***Perron vector*** is the unique vector defined by

$$\mathbf{A}\mathbf{p} = r\mathbf{p}, \qquad \mathbf{p} > \mathbf{0}, \qquad \text{and} \qquad \|\mathbf{p}\|_1 = 1,$$

  and, except for positive multiples of $\mathbf{p}$, there are no other nonnegative eigenvectors for $\mathbf{A}$, regardless of the eigenvalue.
- r is the only eigenvalue on the spectral circle of $\mathbf{A}.$ (8.2.15)
- $r = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x})$ (***the Collatz–Wielandt formula***),

$$\text{where } f(\mathbf{x}) = \min_{\substack{1\le i\le n\\ x_i\neq 0}} \frac{[\mathbf{A}\mathbf{x}]_i}{x_i} \text{ and } \mathcal{N} = \{\mathbf{x} \,|\, \mathbf{x} \geq \mathbf{0} \text{ with } \mathbf{x} \neq \mathbf{0}\}.$$

Note: Our development is the reverse of that of Wielandt and others in the sense that we first proved the existence of the Perron eigenpair $(r, \mathbf{p})$ without reference to $f(\mathbf{x})$, and then we used the Perron eigenpair to established the Collatz-Wielandt formula. Wielandt's approach is to do things the other way around—first prove that $f(\mathbf{x})$ attains a maximum value on $\mathcal{N}$, and then establish existence of the Perron eigenpair by proving that $\max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x}) = \rho(\mathbf{A})$ with the maximum value being attained at a positive eigenvector $\mathbf{p}$.

Exercises for section 8.2

8.2.1. Verify Perron's theorem by by computing the eigenvalues and eigenvectors for

$$\mathbf{A} = \begin{pmatrix} 7 & 2 & 3 \\ 1 & 8 & 3 \\ 1 & 2 & 9 \end{pmatrix}.$$

Find the right-hand Perron vector $\mathbf{p}$ as well as the left-hand Perron vector $\mathbf{q}^T$.

8.2.2. Convince yourself that (8.2.2)–(8.2.5) are indeed true.

8.2.3. Provide the details that explain why the Perron vector is uniquely defined.

8.2.4. Find the Perron root and the Perron vector for

$$\mathbf{A} = \begin{pmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{pmatrix},$$

where $\alpha + \beta = 1$ with $\alpha, \beta > 0$.

8.2.5. Suppose that $\mathbf{A}_{n\times n} > \mathbf{0}$ has $\rho(\mathbf{A}) = r$.

(a) Explain why $\lim_{k\to\infty}(\mathbf{A}/r)^k$ exists.

(b) Explain why $\lim_{k\to\infty}(\mathbf{A}/r)^k = \mathbf{G} > \mathbf{0}$ is the projector onto $N(\mathbf{A} - r\mathbf{I})$ along $R(\mathbf{A} - r\mathbf{I})$.

(c) Explain why $rank\,(\mathbf{G}) = 1$.

8.2.6. Prove that if every row (or column) sum of $\mathbf{A}_{n\times n} > \mathbf{0}$ is equal to ρ, then $\rho\,(\mathbf{A}) = \rho$.

8.2.7. Prove that if $\mathbf{A}_{n\times n} > \mathbf{0}$, then

$$\min_i \sum_{j=1}^{n} a_{ij} \le \rho\,(\mathbf{A}) \le \max_i \sum_{j=1}^{n} a_{ij}.$$

Hint: Recall Example 7.10.2 (p. 619).

8.2.8. To show the extent to which the hypothesis of positivity cannot be relaxed in Perron's theorem, construct examples of square matrices $\mathbf{A}$ such that $\mathbf{A} \ge \mathbf{0}$, but $\mathbf{A} \not> \mathbf{0}$ (i.e., $\mathbf{A}$ has at least one zero entry), with $r = \rho\,(\mathbf{A}) \in \sigma\,(\mathbf{A})$ that demonstrate the validity of the following statements. Different examples may be used for the different statements.

(a) r can be 0.

(b) $alg\ mult_{\mathbf{A}}\,(r)$ can be greater than 1.

(c) $index(r)$ can be greater than 1.

(d) $N(\mathbf{A} - r\mathbf{I})$ need not contain a positive eigenvector.

(e) r need not be the only eigenvalue on the spectral circle.

8.2.9. Establish the min-max version of the Collatz–Wielandt formula that says the Perron root for $\mathbf{A} > \mathbf{0}$ is given by $r = \min_{\mathbf{x}\in\mathcal{P}} g(\mathbf{x})$, where

$$g(\mathbf{x}) = \max_{1\le i\le n} \frac{[\mathbf{A}\mathbf{x}]_i}{x_i} \quad \text{and} \quad \mathcal{P} = \{\mathbf{x} \,|\, \mathbf{x} > \mathbf{0}\}.$$

8.2.10. Notice that $\mathcal{N} = \{\mathbf{x} \,|\, \mathbf{x} \ge \mathbf{0} \text{ with } \mathbf{x} \ne \mathbf{0}\}$ is used in the max-min version of the Collatz–Wielandt formula on p. 666, but $\mathcal{P} = \{\mathbf{x} \,|\, \mathbf{x} > \mathbf{0}\}$ is used in the min-max version in Exercise 8.2.9. Give an example of a matrix $\mathbf{A} > \mathbf{0}$ that shows $r \ne \min_{\mathbf{x}\in\mathcal{N}} g(\mathbf{x})$ when $g(\mathbf{x})$ is defined as

$$g(\mathbf{x}) = \max_{\substack{1\le i\le n \\ x_i\ne 0}} \frac{[\mathbf{A}\mathbf{x}]_i}{x_i}.$$

8.3 NONNEGATIVE MATRICES

Now let zeros creep into the picture and investigate the extent to which Perron's results generalize to nonnegative matrices containing at least one zero entry. The first result along these lines shows how to extend the statements on p. 663 to nonnegative matrices by sacrificing the existence of a positive eigenvector for a nonnegative one.

Nonnegative Eigenpair

For $\mathbf{A}_{n\times n} \geq \mathbf{0}$ with $r = \rho(\mathbf{A})$, the following statements are true.

- $r \in \sigma(\mathbf{A})$, (but $r = 0$ is possible). (8.3.1)
- $\mathbf{A}\mathbf{z} = r\mathbf{z}$ for some $\mathbf{z} \in \mathcal{N} = \{\mathbf{x} \mid \mathbf{x} \geq \mathbf{0} \text{ with } \mathbf{x} \neq \mathbf{0}\}$. (8.3.2)
- $r = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x})$, where $f(\mathbf{x}) = \min\limits_{\substack{1\leq i\leq n\\ x_i\neq 0}} \dfrac{[\mathbf{A}\mathbf{x}]_i}{x_i}$ (8.3.3)

  (i.e., the Collatz–Wielandt formula remains valid).

Proof. Consider the sequence of positive matrices $\mathbf{A}_k = \mathbf{A} + (1/k)\mathbf{E} > \mathbf{0}$, where $\mathbf{E}$ is the matrix of all 1's, and let $r_k > 0$ and $\mathbf{p}_k > \mathbf{0}$ denote the Perron root and Perron vector for $\mathbf{A}_k$, respectively. Observe that $\{\mathbf{p}_k\}_{k=1}^{\infty}$ is a bounded set because it's contained in the unit 1-sphere in $\Re^n$. The Bolzano–Weierstrass theorem states that each bounded sequence in $\Re^n$ has a convergent subsequence. Therefore, $\{\mathbf{p}_k\}_{k=1}^{\infty}$ has convergent subsequence

$$\{\mathbf{p}_{k_i}\}_{i=1}^{\infty} \to \mathbf{z}, \text{ where } \mathbf{z} \geq \mathbf{0} \text{ with } \mathbf{z} \neq \mathbf{0} \text{ (because } \mathbf{p}_{k_i} > \mathbf{0} \text{ and } \|\mathbf{p}_{k_i}\|_1 = 1).$$

Since $\mathbf{A}_1 > \mathbf{A}_2 > \cdots > \mathbf{A}$, the result in Example 7.10.2 (p. 619) guarantees that $r_1 \geq r_2 \geq \cdots \geq r$, so $\{r_k\}_{k=1}^{\infty}$ is a monotonic sequence of positive numbers that is bounded below by r. A standard result from analysis guarantees that

$$\lim_{k\to\infty} r_k = r^\star \text{ exists, and } r^\star \geq r. \quad \text{In particular,} \quad \lim_{i\to\infty} r_{k_i} = r^\star \geq r.$$

But $\lim_{k\to\infty} \mathbf{A}_k = \mathbf{A}$ implies $\lim_{i\to\infty} \mathbf{A}_{k_i} \to \mathbf{A}$, so, by using the easily established fact that the limit of a product is the product of the limits (provided that all limits exist), it's also true that

$$\mathbf{A}\mathbf{z} = \lim_{i\to\infty} \mathbf{A}_{k_i}\mathbf{p}_{k_i} = \lim_{i\to\infty} r_{k_i}\mathbf{p}_{k_i} = r^\star\mathbf{z} \implies r^\star \in \sigma(\mathbf{A}) \implies r^\star \leq r.$$

Consequently, $r^\star = r$, and $\mathbf{A}\mathbf{z} = r\mathbf{z}$ with $\mathbf{z} \geq \mathbf{0}$ and $\mathbf{z} \neq \mathbf{0}$. Thus (8.3.1) and (8.3.2) are proven. To prove (8.3.3), let $\mathbf{q}_k^T > \mathbf{0}$ be the left-hand Perron vector of $\mathbf{A}_k$. For every $\mathbf{x} \in \mathcal{N}$ and $k > 0$ we have $\mathbf{q}_k^T\mathbf{x} > 0$ (by (8.2.2)), and

$$\begin{aligned}\mathbf{0} \leq f(\mathbf{x})\mathbf{x} \leq \mathbf{A}\mathbf{x} \leq \mathbf{A}_k\mathbf{x} \implies f(\mathbf{x})\mathbf{q}_k^T\mathbf{x} \leq \mathbf{q}_k^T\mathbf{A}_k\mathbf{x} = r_k\mathbf{q}_k^T\mathbf{x} \implies f(\mathbf{x}) \leq r_k \\ \implies f(\mathbf{x}) \leq r \quad (\text{because } r_k \to r^* = r).\end{aligned}$$

Since $f(\mathbf{z}) = r$ and $\mathbf{z} \in \mathcal{N}$, it follows that $\max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x}) = r$. ∎

This is as far as Perron's theorem can be generalized to nonnegative matrices without additional hypothesis. For example, $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ shows that properties (8.2.11), (8.2.13), and (8.2.14) on p. 667 do not hold for general nonnegative matrices, and $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ shows that (8.2.15) is also lost. Rather than accepting that the major issues concerning spectral properties of nonnegative matrices had been settled, Frobenius had the insight to look below the surface and see that the problem doesn't stem just from the existence of zero entries, but rather from the *positions* of the zero entries. For example, (8.2.13) and (8.2.14) are false for

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \text{ but they are true for } \widetilde{\mathbf{A}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}. \tag{8.3.4}$$

Frobenius's genius was to see the difference between $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ in terms of reducibility and to relate these ideas to spectral properties of nonnegative matrices. Reducibility and graphs were discussed in Example 4.4.6 (p. 202) and Exercise 4.4.20 (p. 209), but for the sake of continuity they are reviewed below.

Reducibility and Graphs

- $\mathbf{A}_{n\times n}$ is said to be a ***reducible matrix*** when there exists a permutation matrix $\mathbf{P}$ such that

  $$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}, \quad \text{where } \mathbf{X} \text{ and } \mathbf{Z} \text{ are both square.}$$

  Otherwise $\mathbf{A}$ is said to be an ***irreducible matrix.***

- $\mathbf{P}^T\mathbf{A}\mathbf{P}$ is called a ***symmetric permutation*** of $\mathbf{A}$. The effect is to interchange rows in the same way as columns are interchanged.

- The ***graph*** $\mathcal{G}(\mathbf{A})$ of $\mathbf{A}$ is defined to be the directed graph on n nodes $\{N_1, N_2, \ldots, N_n\}$ in which there is a directed edge leading *from* N_i *to* N_j if and only if $a_{ij} \neq 0$.

- $\mathcal{G}(\mathbf{P}^T\mathbf{A}\mathbf{P}) = \mathcal{G}(\mathbf{A})$ whenever $\mathbf{P}$ is a permutation matrix—the effect is simply a relabeling of nodes.

- $\mathcal{G}(\mathbf{A})$ is called ***strongly connected*** if for each pair of nodes (N_i, N_k) there is a sequence of directed edges leading from N_i to N_k.

- $\mathbf{A}$ is an irreducible matrix if and only if $\mathcal{G}(\mathbf{A})$ is strongly connected (see Exercise 4.4.20 on p. 209).

For example, the matrix $\mathbf{A}$ in (8.3.4) is reducible because

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{for} \quad \mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and, as can be seen from Figure 8.3.1, $\mathcal{G}(\mathbf{A})$ is not strongly connected because there is no sequence of paths leading from node 1 to node 2. On the other hand, $\widetilde{\mathbf{A}}$ is irreducible, and as shown in Figure 8.3.1, $\mathcal{G}(\widetilde{\mathbf{A}})$ is strongly connected because each node is accessible from the other.

FIGURE 8.3.1

This discussion suggests that some of Perron's properties given on p. 667 extend to nonnegative matrices when the zeros are in just the right positions to insure irreducibility. To prove that this is in fact the case, the following lemma is needed. It shows how to convert a nonnegative irreducible matrix into a positive matrix in a useful fashion.

Converting Nonnegativity & Irreducibility to Positivity

If $\mathbf{A}_{n\times n} \geq \mathbf{0}$ is irreducible, then $(\mathbf{I}+\mathbf{A})^{n-1} > \mathbf{0}$. (8.3.5)

Proof. Let $a_{ij}^{(k)}$ denote the (i,j)-entry in $\mathbf{A}^k$, and observe that

$$a_{ij}^{(k)} = \sum_{h_1,\ldots,h_{k-1}} a_{ih_1}a_{h_1h_2}\cdots a_{h_{k-1}j} > 0$$

if and only if there exists a set of indicies $h_1, h_2, \ldots, h_{k-1}$ such that

$$a_{ih_1} > 0 \quad \text{and} \quad a_{h_1h_2} > 0 \quad \text{and} \quad \cdots \quad \text{and} \quad a_{h_{k-1}j} > 0.$$

In other words, there is a sequence of k paths $N_i \to N_{h_1} \to N_{h_2} \to \cdots \to N_j$ in $\mathcal{G}(\mathbf{A})$ that lead from node N_i to node N_j if and only if $a_{ij}^{(k)} > 0$. The irreducibility of $\mathbf{A}$ insures that $\mathcal{G}(\mathbf{A})$ is strongly connected, so for any pair of nodes (N_i, N_j) there is a sequence of k paths (with $k < n$) from N_i to N_j. This means that for each position (i,j), there is some $0 \leq k \leq n-1$ such that $a_{ij}^{(k)} > 0$, and this guarantees that for each i and j,

$$\left[(\mathbf{I}+\mathbf{A})^{n-1}\right]_{ij} = \left[\sum_{k=0}^{n-1}\binom{n-1}{k}\mathbf{A}^k\right]_{ij} = \sum_{k=0}^{n-1}\binom{n-1}{k}a_{ij}^{(k)} > \mathbf{0}. \quad \blacksquare$$

With the exception of the Collatz–Wielandt formula, we have seen that $\rho(\mathbf{A}) \in \sigma(\mathbf{A})$ is the only property in the list of Perron properties on p. 667 that extends to nonnegative matrices without additional hypothesis. The next theorem shows how adding irreducibility to nonnegativity recovers the Perron properties (8.2.11), (8.2.13), and (8.2.14).

Perron–Frobenius Theorem

If $\mathbf{A}_{n\times n} \geq \mathbf{0}$ is irreducible, then each of the following is true.

- $r = \rho(\mathbf{A}) \in \sigma(\mathbf{A})$ and $r > 0$. (8.3.6)
- $\textit{alg mult}_{\mathbf{A}}(r) = 1$. (8.3.7)
- There exists an eigenvector $\mathbf{x} > \mathbf{0}$ such that $\mathbf{A}\mathbf{x} = r\mathbf{x}$. (8.3.8)
- The unique vector defined by

$$\mathbf{A}\mathbf{p} = r\mathbf{p}, \qquad \mathbf{p} > \mathbf{0}, \qquad \text{and} \qquad \|\mathbf{p}\|_1 = 1, \tag{8.3.9}$$

  is called the ***Perron vector.*** There are no nonnegative eigenvectors for $\mathbf{A}$ except for positive multiples of $\mathbf{p}$, regardless of the eigenvalue.
- The Collatz–Wielandt formula $r = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x})$,

$$\text{where } f(\mathbf{x}) = \min_{\substack{1\leq i\leq n\\ x_i\neq 0}} \frac{[\mathbf{A}\mathbf{x}]_i}{x_i} \text{ and } \mathcal{N} = \{\mathbf{x} \,|\, \mathbf{x} \geq \mathbf{0} \text{ with } \mathbf{x} \neq \mathbf{0}\}$$

  was established in (8.3.3) for all nonnegative matrices, but it is included here for the sake of completeness.

Proof. We already know from (8.3.2) that $r = \rho(\mathbf{A}) \in \sigma(\mathbf{A})$. To prove that $\textit{alg mult}_{\mathbf{A}}(r) = 1$, let $\mathbf{B} = (\mathbf{I} + \mathbf{A})^{n-1} > \mathbf{0}$ be the matrix in (8.3.5). It follows from (7.9.3) that $\lambda \in \sigma(\mathbf{A})$ if and only if $(1+\lambda)^{n-1} \in \sigma(\mathbf{B})$, and $\textit{alg mult}_{\mathbf{A}}(\lambda) = \textit{alg mult}_{\mathbf{B}}\left((1+\lambda)^{n-1}\right)$. Consequently, if $\mu = \rho(\mathbf{B})$, then

$$\mu = \max_{\lambda\in\sigma(\mathbf{A})} |(1+\lambda)|^{n-1} = \left\{\max_{\lambda\in\sigma(\mathbf{A})} |(1+\lambda)|\right\}^{n-1} = (1+r)^{n-1}$$

because when a circular disk $|z| \leq \rho$ is translated one unit to the right, the point of maximum modulus in the resulting disk $|z+1| \leq \rho$ is $z = 1+\rho$ (it's clear if you draw a picture). Therefore, $\textit{alg mult}_{\mathbf{A}}(r) = 1$; otherwise $\textit{alg mult}_{\mathbf{B}}(\mu) > 1$, which is impossible because $\mathbf{B} > \mathbf{0}$. To see that $\mathbf{A}$ has a positive eigenvector

associated with r, recall from (8.3.2) that there exists a nonnegative eigenvector $\mathbf{x} \geq \mathbf{0}$ associated with r. It's a simple consequence of (7.9.9) that if $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(f(\lambda), \mathbf{x})$ is an eigenpair for $f(\mathbf{A})$ (Exercise 7.9.9, p. 613), so $(r, \mathbf{x})$ being an eigenpair for $\mathbf{A}$ implies that $(\mu, \mathbf{x})$ is an eigenpair for $\mathbf{B}$. Hence (8.2.10) insures that $\mathbf{x}$ must be a positive multiple of the Perron vector of $\mathbf{B}$, and thus $\mathbf{x}$ must in fact be positive. Now, $r > 0$; otherwise $\mathbf{Ax} = \mathbf{0}$, which is impossible because $\mathbf{A} \geq \mathbf{0}$ and $\mathbf{x} > \mathbf{0}$ forces $\mathbf{Ax} > \mathbf{0}$. The argument used to prove (8.2.10) also proves (8.3.9). ∎

Example 8.3.1

Problem: Suppose that $\mathbf{A}_{n\times n} \geq \mathbf{0}$ is irreducible with $r = \rho(\mathbf{A})$, and suppose that $r\mathbf{z} \leq \mathbf{Az}$ for $\mathbf{z} \geq \mathbf{0}$. Explain why $r\mathbf{z} = \mathbf{Az}$, and $\mathbf{z} > \mathbf{0}$.

Solution: If $r\mathbf{z} < \mathbf{Az}$, then by using the Perron vector $\mathbf{q} > \mathbf{0}$ for $\mathbf{A}^T$ we have

$$(\mathbf{A} - r\mathbf{I})\mathbf{z} \geq \mathbf{0} \quad \Longrightarrow \quad \mathbf{q}^T(\mathbf{A} - r\mathbf{I})\mathbf{z} > \mathbf{0},$$

which is impossible since $\mathbf{q}^T(\mathbf{A} - r\mathbf{I}) = \mathbf{0}$. Thus $r\mathbf{z} = \mathbf{Az}$, and since $\mathbf{z}$ must be a multiple of the Perron vector for $\mathbf{A}$ by (8.3.9), we also have that $\mathbf{z} > \mathbf{0}$.

The only property in the list on p. 667 that irreducibility is not able to salvage is (8.2.15), which states that there is only one eigenvalue on the spectral circle. Indeed, $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is nonnegative and irreducible, but the eigenvalues ± 1 are both on the unit circle. The property of having (or not having) only one eigenvalue on the spectral circle divides the set of nonnegative irreducible matrices into two important classes.

Primitive Matrices

- A nonnegative irreducible matrix $\mathbf{A}$ having only one eigenvalue, $r = \rho(\mathbf{A})$, on its spectral circle is said to be a ***primitive matrix.***
- A nonnegative irreducible matrix having $h > 1$ eigenvalues on its spectral circle is called ***imprimitive,*** and h is referred to as ***index of imprimitivity.***
- A nonnegative irreducible matrix $\mathbf{A}$ with $r = \rho(\mathbf{A})$ is primitive if and only if $\lim_{k\to\infty}(\mathbf{A}/r)^k$ exists, in which case

$$\lim_{k\to\infty}\left(\frac{\mathbf{A}}{r}\right)^k = \mathbf{G} = \frac{\mathbf{pq}^T}{\mathbf{q}^T\mathbf{p}} > \mathbf{0}, \tag{8.3.10}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the respective Perron vectors for $\mathbf{A}$ and $\mathbf{A}^T$. $\mathbf{G}$ is the (spectral) projector onto $N(\mathbf{A} - r\mathbf{I})$ along $R(\mathbf{A} - r\mathbf{I})$.

Proof of (8.3.10). The Perron–Frobenius theorem insures that $1 = \rho(\mathbf{A}/r)$ is a simple eigenvalue for $\mathbf{A}/r$, and it's clear that $\mathbf{A}$ is primitive if and only if $\mathbf{A}/r$ is primitive. In other words, $\mathbf{A}$ is primitive if and only if $1 = \rho(\mathbf{A}/r)$ is the only eigenvalue on the unit circle, which is equivalent to saying that $\lim_{k\to\infty}(\mathbf{A}/r)^k$ exists by the results on p. 630. The structure of the limit as described in (8.3.10) is the result of (7.2.12) on p. 518. ∎

The next two results, discovered by Helmut Wielandt (p. 534) in 1950, establish the remarkable fact that the eigenvalues on the spectral circle of an imprimitive matrix are in fact the hth roots of the spectral radius.

Wielandt's Theorem

If $|\mathbf{B}| \le \mathbf{A}_{n\times n}$, where $\mathbf{A}$ is irreducible, then $\rho(\mathbf{B}) \le \rho(\mathbf{A})$. If equality holds (i.e., if $\mu = \rho(\mathbf{A})\,\mathrm{e}^{\mathrm{i}\phi} \in \sigma(\mathbf{B})$ for some ϕ), then

$$\mathbf{B} = \mathrm{e}^{\mathrm{i}\phi}\mathbf{D}\mathbf{A}\mathbf{D}^{-1} \quad \text{for some} \quad \mathbf{D} = \begin{pmatrix} \mathrm{e}^{\mathrm{i}\theta_1} & & & \\ & \mathrm{e}^{\mathrm{i}\theta_2} & & \\ & & \ddots & \\ & & & \mathrm{e}^{\mathrm{i}\theta_n} \end{pmatrix}, \tag{8.3.11}$$

and conversely.

Proof. We already know that $\rho(\mathbf{B}) \le \rho(\mathbf{A})$ by Example 7.10.2 (p. 619). If $\rho(\mathbf{B}) = r = \rho(\mathbf{A})$, and if $(\mu, \mathbf{x})$ is an eigenpair for $\mathbf{B}$ such that $|\mu| = r$, then

$$r|\mathbf{x}| = |\mu|\,|\mathbf{x}| = |\mu\mathbf{x}| = |\mathbf{B}\mathbf{x}| \le |\mathbf{B}|\,|\mathbf{x}| \le \mathbf{A}|\mathbf{x}| \quad \Longrightarrow \quad |\mathbf{B}|\,|\mathbf{x}| = r|\mathbf{x}|$$

because the result in Example 8.3.1 insures that $\mathbf{A}|\mathbf{x}| = r|\mathbf{x}|$, and $|\mathbf{x}| > \mathbf{0}$. Consequently, $(\mathbf{A} - |\mathbf{B}|)|\mathbf{x}| = \mathbf{0}$. But $\mathbf{A} - |\mathbf{B}| \ge \mathbf{0}$, and $|\mathbf{x}| > \mathbf{0}$, so $\mathbf{A} = |\mathbf{B}|$ by (8.2.4). Since $x_k/|x_k|$ is on the unit circle, $x_k/|x_k| = \mathrm{e}^{\mathrm{i}\theta_k}$ for some θ_k. Set

$$\mathbf{D} = \begin{pmatrix} \mathrm{e}^{\mathrm{i}\theta_1} & & & \\ & \mathrm{e}^{\mathrm{i}\theta_2} & & \\ & & \ddots & \\ & & & \mathrm{e}^{\mathrm{i}\theta_n} \end{pmatrix}, \quad \text{and notice that} \quad \mathbf{x} = \mathbf{D}|\mathbf{x}|.$$

Since $|\mu| = r$, there is a $\phi \in \Re$ such that $\mu = r\mathrm{e}^{\mathrm{i}\phi}$, and hence

$$\mathbf{B}\mathbf{D}|\mathbf{x}| = \mathbf{B}\mathbf{x} = \mu\mathbf{x} = r\mathrm{e}^{\mathrm{i}\phi}\mathbf{x} = r\mathrm{e}^{\mathrm{i}\phi}\mathbf{D}|\mathbf{x}| \;\Rightarrow\; \mathrm{e}^{-\mathrm{i}\phi}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}|\mathbf{x}| = r|\mathbf{x}| = \mathbf{A}|\mathbf{x}|. \tag{8.3.12}$$

For convenience, let $\mathbf{C} = \mathrm{e}^{-\mathrm{i}\phi}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}$, and note that $|\mathbf{C}| = |\mathbf{B}| = \mathbf{A}$ to write (8.3.12) as $\mathbf{0} = (|\mathbf{C}| - \mathbf{C})|\mathbf{x}|$. Considering only the real part of this equation

yields $\mathbf{0} = \big(|\mathbf{C}| - \mathrm{Re}\,(\mathbf{C})\big)|\mathbf{x}|$. But $|\mathbf{C}| \geq \mathrm{Re}\,(\mathbf{C})$, and $|\mathbf{x}| > \mathbf{0}$, so it follows from (8.2.4) that $\mathrm{Re}\,(\mathbf{C}) = |\mathbf{C}|$, and hence

$$\mathrm{Re}\,(c_{ij}) = |c_{ij}| = \sqrt{\mathrm{Re}\,(c_{ij})^2 + \mathrm{Im}\,(c_{ij})^2} \implies \mathrm{Im}\,(c_{ij}) = 0 \implies \mathrm{Im}\,(\mathbf{C}) = \mathbf{0}.$$

Therefore, $\mathbf{C} = \mathrm{Re}\,(\mathbf{C}) = |\mathbf{C}| = \mathbf{A}$, which implies $\mathbf{B} = \mathrm{e}^{\mathrm{i}\phi}\mathbf{D}\mathbf{A}\mathbf{D}^{-1}$. Conversely, if $\mathbf{B} = \mathrm{e}^{\mathrm{i}\phi}\mathbf{D}\mathbf{A}\mathbf{D}^{-1}$, then similarity insures that $\rho\,(\mathbf{B}) = \rho\left(\mathrm{e}^{\mathrm{i}\phi}\mathbf{A}\right) = \rho\,(\mathbf{A})$. ∎

h^{th} Roots of $\rho\,(\mathbf{A})$ on Spectral Circle

If $\mathbf{A}_{n\times n} \geq \mathbf{0}$ is irreducible and has h eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_h\}$ on its spectral circle, then each of the following statements is true.

- $\textit{alg mult}_{\mathbf{A}}\,(\lambda_k) = 1$ for $k = 1, 2, \ldots, h$. (8.3.13)
- $\{\lambda_1, \lambda_2, \ldots, \lambda_h\}$ are the hth roots of $r = \rho\,(\mathbf{A})$ given by

$$\{r,\, r\omega,\, r\omega^2,\, \ldots,\, r\omega^{h-1}\}, \quad \text{where} \quad \omega = \mathrm{e}^{2\pi\mathrm{i}/h}. \tag{8.3.14}$$

Proof. Let $\mathcal{S} = \{r, r\mathrm{e}^{\mathrm{i}\theta_1}, \ldots, r\mathrm{e}^{\mathrm{i}\theta_{h-1}}\}$ denote the eigenvalues on the spectral circle of $\mathbf{A}$. Applying (8.3.11) with $\mathbf{B} = \mathbf{A}$ and $\mu = r\mathrm{e}^{\mathrm{i}\theta_k}$ insures the existence of a diagonal matrix $\mathbf{D}_k$ such that $\mathbf{A} = \mathrm{e}^{\mathrm{i}\theta_k}\mathbf{D}_k\mathbf{A}\mathbf{D}_k^{-1}$, thus showing that $\mathrm{e}^{\mathrm{i}\theta_k}\mathbf{A}$ is similar to $\mathbf{A}$. Since r is a simple eigenvalue of $\mathbf{A}$ (by the Perron–Frobenius theorem), $r\mathrm{e}^{\mathrm{i}\theta_k}$ must be a simple eigenvalue of $\mathrm{e}^{\mathrm{i}\theta_k}\mathbf{A}$. But similarity transformations preserve eigenvalues and algebraic multiplicities (because the Jordan structure is preserved), so $r\mathrm{e}^{\mathrm{i}\theta_k}$ must be a simple eigenvalue of $\mathbf{A}$, thus establishing (8.3.13). To prove (8.3.14), consider another eigenvalue $r\mathrm{e}^{\mathrm{i}\theta_s} \in \mathcal{S}$. Again, we can write $\mathbf{A} = \mathrm{e}^{\mathrm{i}\theta_s}\mathbf{D}_s\mathbf{A}\mathbf{D}_s^{-1}$ for some $\mathbf{D}_s$, so

$$\mathbf{A} = \mathrm{e}^{\mathrm{i}\theta_k}\mathbf{D}_k\mathbf{A}\mathbf{D}_k^{-1} = \mathrm{e}^{\mathrm{i}\theta_k}\mathbf{D}_k(\mathrm{e}^{\mathrm{i}\theta_s}\mathbf{D}_s\mathbf{A}\mathbf{D}_s^{-1})\mathbf{D}_k^{-1} = \mathrm{e}^{\mathrm{i}(\theta_k+\theta_r)}(\mathbf{D}_k\mathbf{D}_s)\mathbf{A}(\mathbf{D}_k\mathbf{D}_s)^{-1}$$

and, consequently, $r\mathrm{e}^{\mathrm{i}(\theta_k+\theta_r)}$ is also an eigenvalue on the spectral circle of $\mathbf{A}$. In other words, $\mathcal{S} = \{r, r\mathrm{e}^{\mathrm{i}\theta_1}, \ldots, r\mathrm{e}^{\mathrm{i}\theta_{h-1}}\}$ is closed under multiplication. This means that $\mathcal{G} = \{1, \mathrm{e}^{\mathrm{i}\theta_1}, \ldots, \mathrm{e}^{\mathrm{i}\theta_{h-1}}\}$ is closed under multiplication, and it follows that $\mathcal{G}$ is a finite commutative group of order h. A standard result from algebra states that the hth power of every element in a finite group of order h must be the identity element in the group. Therefore, $(\mathrm{e}^{\mathrm{i}\theta_k})^h = 1$ for each $0 \leq k \leq h-1$, so $\mathcal{G}$ is the set of the hth roots of unity $\mathrm{e}^{2\pi k\mathrm{i}/h}$ $(0 \leq k \leq h-1)$, and thus $\mathcal{S}$ must be the hth roots of r. ∎

Combining the preceding results reveals just how special the spectrum of an imprimitive matrix is.

Rotational Invariance

If $\mathbf{A}$ is imprimitive with h eigenvalues on its spectral circle, then $\sigma(\mathbf{A})$ is invariant under rotation about the origin through an angle $2\pi/h$. No rotation less than $2\pi/h$ can preserve $\sigma(\mathbf{A})$. (8.3.15)

Proof. Since $\lambda \in \sigma(\mathbf{A}) \iff \lambda e^{2\pi i/h} \in \sigma(e^{2\pi i/h}\mathbf{A})$, it follows that $\sigma(e^{2\pi i/h}\mathbf{A})$ is $\sigma(\mathbf{A})$ rotated through $2\pi/h$. But (8.3.11) and (8.3.14) insure that $\mathbf{A}$ and $e^{2\pi i/h}\mathbf{A}$ are similar and, consequently, $\sigma(\mathbf{A}) = \sigma(e^{2\pi i/h}\mathbf{A})$. No rotation less than $2\pi/h$ can keep $\sigma(\mathbf{A})$ invariant because (8.3.14) makes it clear that the eigenvalues on the spectral circle won't go back into themselves for rotations less than $2\pi/h$. ∎

Example 8.3.2

The Spectral Projector Is Positive. We already know from (8.3.10) that if $\mathbf{A}$ is a primitive matrix, and if $\mathbf{G}$ is the spectral projector associated with $r = \rho(\mathbf{A})$, then $\mathbf{G} > \mathbf{0}$.

Problem: Explain why this is also true for an imprimitive matrix. In other words, establish the fact that *if* $\mathbf{G}$ *is the spectral projector associated with* $r = \rho(\mathbf{A})$ *for any nonnegative irreducible matrix* $\mathbf{A}$, *then* $\mathbf{G} > \mathbf{0}$.

Solution: Being imprimitive means that $\mathbf{A}$ is nonnegative and irreducible with more than one eigenvalue on the spectral circle. However, (8.3.13) says that each eigenvalue on the spectral circle is simple, so the results concerning Cesàro summability on p. 633 can be applied to $\mathbf{A}/r$ to conclude that

$$\lim_{k\to\infty} \frac{\mathbf{I} + (\mathbf{A}/r) + \cdots + (\mathbf{A}/r)^{k-1}}{k} = \mathbf{G},$$

where $\mathbf{G}$ is the spectral projector onto $N((\mathbf{A}/r) - \mathbf{I}) = N(\mathbf{A} - r\mathbf{I})$ along $R((\mathbf{A}/r) - \mathbf{I}) = R(\mathbf{A} - r\mathbf{I})$. Since r is a simple eigenvalue the same argument used to establish (8.3.10) (namely, invoking (7.2.12) on p. 518) shows that

$$\mathbf{G} = \frac{\mathbf{p}\mathbf{q}^T}{\mathbf{q}^T\mathbf{p}} > \mathbf{0},$$

where $\mathbf{p}$ and $\mathbf{q}$ are the respective Perron vectors for $\mathbf{A}$ and $\mathbf{A}^T$.

Trying to determine if an irreducible matrix $\mathbf{A} \geq \mathbf{0}$ is primitive or imprimitive by finding the eigenvalues is generally a difficult task, so it's natural to ask if there's another way. It turns out that there is, and, as the following example shows, determining primitivity can sometimes be trivial.

Example 8.3.3

Sufficient Condition for Primitivity. If a nonnegative irreducible matrix $\mathbf{A}$ has at least one positive diagonal element, then $\mathbf{A}$ is primitive.

Proof. Suppose there are $h > 1$ eigenvalues on the spectral circle. We know from (8.3.15) that if $\lambda_0 \in \sigma(\mathbf{A})$, then $\lambda_k = \lambda_0 e^{2\pi i k/h} \in \sigma(\mathbf{A})$ for $k = 0, 1, \ldots, h-1$, so

$$\sum_{k=0}^{h-1} \lambda_k = \lambda_0 \sum_{k=0}^{h-1} e^{2\pi i k/h} = 0 \quad \text{(roots of unity sum to 1—see p. 357).}$$

This implies that the sum of *all* of the eigenvalues is zero. In other words,

- if $\mathbf{A}$ is imprimitive, then $trace(\mathbf{A}) = 0$. (Recall (7.1.7) on p. 494.)

Therefore, if $\mathbf{A}$ has a positive diagonal entry, then $\mathbf{A}$ must be primitive.

Another of Frobenius's contributions was to show how the powers of a nonnegative matrix determine whether or not the matrix is primitive. The exact statement is as follows.

Frobenius's Test for Primitivity

$\mathbf{A} \geq \mathbf{0}$ is primitive if and only if $\mathbf{A}^m > \mathbf{0}$ for some $m > 0$. (8.3.16)

Proof. First assume that $\mathbf{A}^m > \mathbf{0}$ for some m. This implies that $\mathbf{A}$ is irreducible; otherwise there exists a permutation matrix such that

$$\mathbf{A} = \mathbf{P}\begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}\mathbf{P}^T \implies \mathbf{A}^m = \mathbf{P}\begin{pmatrix} \mathbf{X}^m & \star \\ \mathbf{0} & \mathbf{Z}^m \end{pmatrix}\mathbf{P}^T \text{ has zero entries.}$$

Suppose that $\mathbf{A}$ has h eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_h\}$ on its spectral circle so that $r = \rho(\mathbf{A}) = |\lambda_1| = \cdots = |\lambda_h| > |\lambda_{h+1}| > \cdots > |\lambda_n|$. Since $\lambda \in \sigma(\mathbf{A})$ implies $\lambda^m \in \sigma(\mathbf{A}^m)$ with $alg\ mult_{\mathbf{A}}(\lambda) = alg\ mult_{\mathbf{A}^m}(\lambda^m)$ (consider the Jordan form—Exercise 7.9.9 on p. 613), it follows that λ_k^m $(1 \leq k \leq h)$ is on the spectral circle of $\mathbf{A}^m$ with $alg\ mult_{\mathbf{A}}(\lambda_k) = alg\ mult_{\mathbf{A}^m}(\lambda_k^m)$. Perron's theorem (p. 667) insures that $\mathbf{A}^m$ has only one eigenvalue (which must be r^m) on its spectral circle, so $r^m = \lambda_1^m = \lambda_2^m = \cdots = \lambda_h^m$. But this means that

$$alg\ mult_{\mathbf{A}}(r) = alg\ mult_{\mathbf{A}^m}(r^m) = h,$$

and therefore $h = 1$ by (8.3.7). Conversely, if $\mathbf{A}$ is primitive with $r = \rho(\mathbf{A})$, then $\lim_{k\to\infty}(\mathbf{A}/r)^k > \mathbf{0}$ by (8.3.10). Hence there must be some m such that $(\mathbf{A}/r)^m > \mathbf{0}$, and thus $\mathbf{A}^m > \mathbf{0}$. ∎

Example 8.3.4

Suppose that we wish to decide whether or not a nonnegative matrix $\mathbf{A}$ is primitive by computing the sequence of powers $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^3, \ldots$. Since this can be a laborious task, it would be nice to know when we have computed enough powers of $\mathbf{A}$ to render a judgement. Unfortunately there is nothing in the statement or proof of Frobenius's test to help us with this decision. But Wielandt provided an answer by proving that a nonnegative matrix $\mathbf{A}_{n\times n}$ is primitive if and only if $\mathbf{A}^{n^2-2n+2} > \mathbf{0}$. Furthermore, $n^2 - 2n + 2$ is the smallest such exponent that works for the class of $n \times n$ primitive matrices having all zeros on the diagonal—see Exercise 8.3.9.

Problem: Determine whether or not $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 3 & 4 & 0 \end{pmatrix}$ is primitive.

Solution: Since $\mathbf{A}$ has zeros on the diagonal, the result in Example 8.3.3 doesn't apply, so we are forced into computing powers of $\mathbf{A}$. This job is simplified by noticing that if $\mathbf{B} = \beta(\mathbf{A})$ is the Boolean matrix that results from setting

$$b_{ij} = \begin{cases} 1 & \text{if } a_{ij} > 0, \\ 0 & \text{if } a_{ij} = 0, \end{cases}$$

then $[\mathbf{B}^k]_{ij} > 0$ if and only if $[\mathbf{A}^k]_{ij} > 0$ for every $k > 0$. This means that instead of using $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^3, \ldots$ to decide on primitivity, we need only compute

$$\mathbf{B}_1 = \beta(\mathbf{A}), \quad \mathbf{B}_2 = \beta(\mathbf{B}_1\mathbf{B}_1), \quad \mathbf{B}_3 = \beta(\mathbf{B}_1\mathbf{B}_2), \quad \mathbf{B}_4 = \beta(\mathbf{B}_1\mathbf{B}_3), \ldots,$$

going no further than $\mathbf{B}_{n^2-2n+2}$, and these computations require only Boolean operations **AND** and **OR**. The matrix $\mathbf{A}$ in this example is primitive because

$$\mathbf{B}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \ \mathbf{B}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \ \mathbf{B}_3 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \ \mathbf{B}_4 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \ \mathbf{B}_5 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

The powers of an irreducible matrix $\mathbf{A} \geq \mathbf{0}$ can tell us if $\mathbf{A}$ has more than one eigenvalue on its spectral circle, but the powers of $\mathbf{A}$ provide no clue to the number of such eigenvalues. The next theorem shows how the index of imprimitivity can be determined without explicitly calculating the eigenvalues.

Index of Imprimitivity

If $c(x) = x^n + c_{k_1}x^{n-k_1} + c_{k_2}x^{n-k_2} + \cdots + c_{k_s}x^{n-k_s} = 0$ is the characteristic equation of an imprimitive matrix $\mathbf{A}_{n\times n}$ in which only the terms with nonzero coefficients are listed (i.e., each $c_{k_j} \neq 0$, and $n > (n - k_1) > \cdots > (n - k_s)$), then the index of imprimitivity h is the greatest common divisor of $\{k_1, k_2, \ldots, k_s\}$.

Proof. We know from (8.3.15) that if $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues of $\mathbf{A}$ (including multiplicities), then $\{\omega\lambda_1, \omega\lambda_2, \ldots, \omega\lambda_n\}$ are also the eigenvalues of $\mathbf{A}$, where $\omega = \mathrm{e}^{2\pi \mathrm{i}/h}$. It follows from the results on p. 494 that

$$c_{k_j} = (-1)^{k_j} \sum_{1 \le i_1 < \cdots < i_{k_j} \le n} \lambda_{i_1} \cdots \lambda_{i_{k_j}} = (-1)^{k_j} \sum_{1 \le i_1 < \cdots < i_{k_j} \le n} \omega\lambda_{i_1} \cdots \omega\lambda_{i_{k_j}} = \omega^{k_j} c_{k_j} \implies \omega^{k_j} = 1.$$

Therefore, h must divide each k_j. If d divides each k_j with $d > h$, then $\gamma^{-k_j} = 1$ for $\gamma = \mathrm{e}^{2\pi \mathrm{i}/d}$. Hence $\gamma\lambda \in \sigma(\mathbf{A})$ for each $\lambda \in \sigma(\mathbf{A})$ because $c(\gamma\lambda) = 0$. But this means that $\sigma(\mathbf{A})$ is invariant under a rotation through an angle $(2\pi/d) < (2\pi/h)$, which, by (8.3.15), is impossible. ∎

Example 8.3.5

Problem: Find the index of imprimitivity of $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 0 \end{pmatrix}$.

Solution: Using the principal minors to compute the characteristic equation as illustrated in Example 7.1.2 (p. 496) produces the characteristic equation

$$c(x) = x^4 - 5x^2 + 4 = 0,$$

so that $k_1 = 2$ and $k_2 = 4$. Since $\gcd\{2, 4\} = 2$, it follows that $h = 2$. The characteristic equation is relatively simple in this example, so the eigenvalues can be explicitly determined to be $\{\pm 2, \pm 1\}$. This corroborates the fact that $h = 2$. Notice also that this illustrates the property that $\sigma(\mathbf{A})$ is invariant under rotation through an angle $2\pi/h = \pi$.

More is known about nonnegative matrices than what has been presented here—in fact, there are entire books on the subject. But before moving on to applications, there is a result that Frobenius discovered in 1912 that is worth mentioning because it completely reveals the structure of an imprimitive matrix.

Frobenius Form

For each imprimitive matrix $\mathbf{A}$ with index of imprimitivity $h > 1$, there exists a permutation matrix $\mathbf{P}$ such that

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_{12} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{23} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{h-1,h} \\ \mathbf{A}_{h1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where the zero blocks on the main diagonal are square.

Example 8.3.6

Leontief's[91] Input–Output Economic Model. Suppose that n major industries in a closed economic system each make one commodity, and let a *J-unit* be what industry J produces that sells for \$1. For example, the Boeing Company makes airplanes, and the Champion Company makes rivets, so a *BOEING-unit* is only a tiny fraction of an airplane, but a *CHAMPION-unit* might be several rivets. If

$$0 \leq s_j = \#\ \textit{J-units} \text{ produced by industry } J \text{ each year, and if}$$
$$0 \leq a_{ij} = \#\ \textit{I-units} \text{ needed to produce one } \textit{J-unit},$$

then

$$a_{ij}s_j = \#\ \textit{I-units} \text{ consumed by industry } J \text{ each year, and}$$
$$\sum_{j=1}^{n} a_{ij}s_j = \#\ \textit{I-units} \text{ consumed by all industries each year,}$$

so

$$d_i = s_i - \sum_{j=1}^{n} a_{ij}s_j = \#\ \textit{I-units} \text{ available to the public (nonindustry) each year.}$$

Consider $\mathbf{d} = (d_1, d_2, \ldots, d_n)^T$ to be the public ***demand vector,*** and think of $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ as the industrial ***supply vector.***

Problem: Determine the supply $\mathbf{s} \geq \mathbf{0}$ that is required to satisfy a given demand $\mathbf{d} \geq \mathbf{0}$.

Solution: At first glance the problem seems to be trivial because the equations $d_i = s_i - \sum_{j=1}^{n} a_{ij}s_j$ translate to $(\mathbf{I}-\mathbf{A})\mathbf{s} = \mathbf{d}$, so if $\mathbf{I}-\mathbf{A}$ is nonsingular, then $\mathbf{s} = (\mathbf{I}-\mathbf{A})^{-1}\mathbf{d}$. The catch is that this solution may have negative components in spite of the fact that $\mathbf{A} \geq \mathbf{0}$. So something must be added. It's not unreasonable to assume that major industries are *strongly connected* in the sense that the commodity of each industry is either directly or indirectly needed to produce all commodities in the system. In other words, it's reasonable to assume that

[91] Wassily Leontief (1906–1999) was the 1973 Nobel Laureate in Economics. He was born in St. Petersburg (now Leningrad), where his father was a professor of economics. After receiving his undergraduate degree in economics at the University of Leningrad in 1925, Leontief went to the University of Berlin to earn a Ph.D. degree. He migrated to New York in 1931 and moved to Harvard University in 1932, where he became Professor of Economics in 1946. Leontief spent a significant portion of his career developing and applying his input–output analysis, which eventually led to the famous "Leontief paradox." In the U.S. economy of the 1950s, labor was considered to be scarce while capital was presumed to be abundant, so the prevailing thought was that U.S. foreign trade was predicated on trading capital-intensive goods for labor-intensive goods. But Leontief's input–output tables revealed that just the opposite was true, and this contributed to his fame. One of Leontief's secret weapons was the computer. He made use of large-scale computing techniques (relative to the technology of the 1940s and 1950s), and he was among the first to put the Mark I (one of the first electronic computers) to work on nonmilitary projects in 1943.

$\mathcal{G}(\mathbf{A})$ is a strongly connected graph so that in addition to being nonnegative, $\mathbf{A}$ is an *irreducible* matrix. Furthermore, it's not unreasonable to assume that $\rho(\mathbf{A}) < 1$. To understand why, notice that the jth column sum of $\mathbf{A}$ is

$$\begin{aligned} c_j = \sum_{i=1}^{n} a_{ij} &= \text{total number of all units required to make one } J\text{-}unit \\ &= \text{total number of dollars spent by } J \text{ to create \$1 of revenue.} \end{aligned}$$

In a healthy economy all major industries should have $c_j \leq 1$, and there should be at least one major industry such that $c_j < 1$. This means that there exists a matrix $\mathbf{E} \geq \mathbf{0}$, but $\mathbf{E} \neq \mathbf{0}$, such that each column sum of $\mathbf{A} + \mathbf{E}$ is 1, so

$$\mathbf{e}^T(\mathbf{A} + \mathbf{E}) = \mathbf{e}^T, \quad \text{where} \quad \mathbf{e}^T \text{ is the row of all 1's.}$$

This forces $\rho(\mathbf{A}) < 1$; otherwise the Perron vector $\mathbf{p} > \mathbf{0}$ for $\mathbf{A}$ can be used to write

$$1 = \mathbf{e}^T\mathbf{p} = \mathbf{e}^T(\mathbf{A} + \mathbf{E})\mathbf{p} = 1 + \mathbf{e}^T\mathbf{E}\mathbf{p} > 1$$

because

$$\mathbf{E} \geq \mathbf{0},\ \mathbf{E} \neq \mathbf{0},\ \mathbf{p} > \mathbf{0} \implies \mathbf{E}\mathbf{p} > \mathbf{0}.$$

(Conditions weaker than the column-sum condition can also force $\rho(\mathbf{A}) < 1$—see Example 7.10.3 on p. 620.) The assumption that $\mathbf{A}$ is a nonnegative irreducible matrix whose spectral radius is $\rho(\mathbf{A}) < 1$ combined with the Neumann series (p. 618) provides the conclusion that

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k > \mathbf{0}.$$

Positivity is guaranteed by the irreducibility of $\mathbf{A}$ because the same argument given on p. 672 that is to prove (8.3.5) also applies here. Therefore, for each demand vector $\mathbf{d} \geq \mathbf{0}$, there exists a unique supply vector given by $\mathbf{s} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{d}$, which is necessarily positive. The fact that $(\mathbf{I} - \mathbf{A})^{-1} > \mathbf{0}$ and $\mathbf{s} > \mathbf{0}$ leads to the interesting conclusion that an increase in public demand by just *one* unit from a *single* industry will force an increase in the output of *all* industries.

Note: The matrix $\mathbf{I} - \mathbf{A}$ is an M-matrix as defined and discussed in Example 7.10.7 (p. 626). The realization that M-matrices are naturally present in economic models provided some of the motivation for studying M-matrices during the first half of the twentieth century. Some of the M-matrix properties listed on p. 626 were independently discovered and formulated in economic terms.

Example 8.3.7

Leslie Population Age Distribution Model. Divide a population of females into age groups $G_1, G_2, \ldots, G_n$, where each group covers the same number of years. For example,

$$\begin{aligned} G_1 &= \text{all females under age 10,} \\ G_2 &= \text{all females from age 10 up to 20,} \\ G_1 &= \text{all females from age 20 up to 30,} \\ &\vdots \end{aligned}$$

Consider discrete points in time, say $t = 0, 1, 2, \ldots$ years, and let b_k and s_k denote the birth rate and survival rate for females in G_k. That is, let

$$\begin{aligned} b_k &= \text{Expected number of daughters produced by a female in } G_k, \\ s_k &= \text{Proportion of females in } G_k \text{ at time } t \text{ that are in } G_{k+1} \text{ at time } t+1. \end{aligned}$$

If

$$f_k(t) = \text{Number of females in } G_k \text{ at time } t,$$

then it follows that

$$f_1(t+1) = f_1(t)b_1 + f_2(t)b_2 + \cdots + f_n(t)b_n$$

and (8.3.17)

$$f_k(t+1) = f_{k-1}(t)s_{k-1} \quad \text{for } k = 2, 3, \ldots, n.$$

Furthermore,

$$F_k(t) = \frac{f_k(t)}{f_1(t) + f_2(t) + \cdots + f_n(t)} = \%\text{ of population in } G_k \text{ at time } t.$$

The vector $\mathbf{F}(t) = (F_1(t), F_2(t), \ldots, F_n(t))^T$ represents the *population age distribution* at time t, and, provided that it exists, $\mathbf{F}^\star = \lim_{t\to\infty} \mathbf{F}(t)$ is the *long-run* (*or steady-state*) *age distribution.*

Problem: Assuming that $s_1, \ldots, s_n$ and $b_2, \ldots, b_n$ are positive, explain why the population age distribution approaches a steady state, and then describe it. In other words, show that $\mathbf{F}^\star = \lim_{t\to\infty} \mathbf{F}(t)$ exists, and determine its value.

Solution: The equations in (8.3.17) constitute a system of homogeneous difference equations that can be written in matrix form as

$$\mathbf{f}(t+1) = \mathbf{L}\mathbf{f}(t), \quad \text{where} \quad \mathbf{L} = \begin{pmatrix} b_1 & b_2 & \cdots & b_{n-1} & b_n \\ s_1 & 0 & \cdots & \cdots & 0 \\ 0 & s_2 & 0 & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}_{n\times n}. \tag{8.3.18}$$

The matrix $\mathbf{L}$ is called the ***Leslie matrix*** in honor of P. H. Leslie who used this model in 1945. Notice that in addition to being nonnegative, $\mathbf{L}$ is also irreducible when $s_1, \ldots, s_n$ and $b_2, \ldots, b_n$ are positive because the graph $\mathcal{G}(\mathbf{L})$ is strongly connected. Moreover, $\mathbf{L}$ is primitive. This is obvious if in addition to $s_1, \ldots, s_n$ and $b_2, \ldots, b_n$ being positive we have $b_1 > 0$ (recall Example 8.3.3 on p. 678). But even if $b_1 = 0$, $\mathbf{L}$ is still primitive because $\mathbf{L}^{n+2} > \mathbf{0}$ (recall (8.3.16) on p. 678). The technique on p. 679 also can be used to show primitivity (Exercise 8.3.11). Consequently, (8.3.10) on p. 674 guarantees that

$$\lim_{t\to\infty} \left(\frac{\mathbf{L}}{r}\right)^t = \mathbf{G} = \frac{\mathbf{p}\mathbf{q}^T}{\mathbf{q}^T\mathbf{p}} > \mathbf{0},$$

where $\mathbf{p} > \mathbf{0}$ and $\mathbf{q} > \mathbf{0}$ are the respective Perron vectors for $\mathbf{L}$ and $\mathbf{L}^T$. If we combine this with the fact that the solution to the system of difference equations in (8.3.18) is $\mathbf{f}(t) = \mathbf{L}^t\mathbf{f}(0)$ (p. 617), and if we assume that $\mathbf{f}(0) \neq \mathbf{0}$, then we arrive at the conclusion that

$$\lim_{t\to\infty} \frac{\mathbf{f}(t)}{r^t} = \mathbf{G}\mathbf{f}(0) = \mathbf{p}\left(\frac{\mathbf{q}^T\mathbf{f}(0)}{\mathbf{q}^T\mathbf{p}}\right) \text{ and } \lim_{t\to\infty} \left\|\frac{\mathbf{f}(t)}{r^t}\right\|_1 = \frac{\mathbf{q}^T\mathbf{f}(0)}{\mathbf{q}^T\mathbf{p}} > 0 \tag{8.3.19}$$

(because $\|\star\|_1$ is a continuous function—Exercise 5.1.7 on p. 277). Now

$$F_k(t) = \frac{f_k(t)}{\|\mathbf{f}(t)\|_1} = \text{ \% of population that is in } G_k \text{ at time } t$$

is the quantity of interest, and (8.3.19) allows us to conclude that

$$\begin{aligned}\mathbf{F}^\star = \lim_{t\to\infty} \mathbf{F}(t) &= \lim_{t\to\infty} \frac{\mathbf{f}(t)}{\|\mathbf{f}(t)\|_1} = \lim_{t\to\infty} \frac{\mathbf{f}(t)/r^t}{\|\mathbf{f}(t)\|_1/r^t} \\ &= \frac{\lim_{t\to\infty} \mathbf{f}(t)/r^t}{\lim_{t\to\infty} \|\mathbf{f}(t)\|_1/r^t} = \mathbf{p} \quad \text{(the Perron vector!).}\end{aligned}$$

In other words, while the numbers in the various age groups may increase or decrease, depending on the value of r (Exercise 8.3.10), the proportion of individuals in each age group becomes stable as time increases. And because the steady-state age distribution is given by the Perron vector of $\mathbf{L}$, each age group must eventually contain a positive fraction of the population.

Exercises for section 8.3

8.3.1. Let $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 3 & 0 & 3 \\ 0 & 2 & 0 \end{pmatrix}$.

(a) Show that $\mathbf{A}$ is irreducible.
(b) Find the Perron root and Perron vector for $\mathbf{A}$.
(c) Find the number of eigenvalues on the spectral circle of $\mathbf{A}$.

8.3.2. Suppose that the index of imprimitivity of a 5×5 nonnegative irreducible matrix $\mathbf{A}$ is $h = 3$. Explain why $\mathbf{A}$ must be singular with $alg\ mult_{\mathbf{A}}(0) = 2$.

8.3.3. Suppose that $\mathbf{A}$ is a nonnegative matrix that possesses a positive spectral radius and a corresponding positive eigenvector. Does this force $\mathbf{A}$ to be irreducible?

8.3.4. Without computing the eigenvalues or the characteristic polynomial, explain why $\sigma(\mathbf{P}_n) = \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$, where $\omega = e^{2\pi i/n}$ for

$$\mathbf{P}_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

8.3.5. Determine whether $\mathbf{A} = \begin{pmatrix} 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 0 & 9 & 2 & 0 & 4 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$ is reducible or irreducible.

8.3.6. Determine whether the matrix $\mathbf{A}$ in Exercise 8.3.5 is primitive or imprimitive.

8.3.7. A matrix $\mathbf{S}_{n\times n} \geq \mathbf{0}$ having row sums less than or equal to 1 with at least one row sum less than 1 is called a ***substochastic matrix.***

(a) Explain why $\rho(\mathbf{S}) \leq 1$ for every substochastic matrix.

(b) Prove that $\rho(\mathbf{S}) < 1$ for every *irreducible* substochastic matrix.

8.3.8. A nonnegative matrix for which each row sum is 1 is called a ***stochastic matrix*** (some say *row*-stochastic). Prove that if $\mathbf{A}_{n\times n}$ is nonnegative and irreducible with $r = \rho(\mathbf{A})$, then $\mathbf{A}$ is similar to $r\mathbf{P}$ for some irreducible stochastic matrix $\mathbf{P}$. **Hint:** Consider $\mathbf{D} = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{pmatrix}$, where the p_k's are the components of the Perron vector for $\mathbf{A}$.

8.3.9. Wielandt constructed the matrix $\mathbf{W}_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 \end{pmatrix}$ to show that $\mathbf{W}^{n^2-2n+2} > \mathbf{0}$, but $[\mathbf{W}^{n^2-2n+1}]_{11} = 0$. Verify that this is true for $n = 4$.

8.3.10. In the Leslie population model on p. 683, explain what happens to the vector $\mathbf{f}(t)$ as $t \to \infty$ depending on whether $r < 1$, $r = 1$, or $r > 1$.

8.3.11. Use the characteristic equation as described on p. 679 to show that the Leslie matrix in (8.3.18) is primitive even if $b_1 = 0$ (assuming all other b_k 's and s_k 's are positive).

8.3.12. A matrix $\mathbf{A} \in \Re^{n \times n}$ is said to be ***essentially positive*** if $\mathbf{A}$ is irreducible and $a_{ij} \geq 0$ for every $i \neq j$. Prove that each of the following statements is equivalent to saying that $\mathbf{A}$ is essentially positive.

(a) There exists some $\alpha \in \Re$ such that $\mathbf{A} + \alpha \mathbf{I}$ is primitive.

(b) $e^{t\mathbf{A}} > \mathbf{0}$ for all $t > 0$.

8.3.13. Let $\mathbf{A}$ be an essentially positive matrix as defined in Exercise 8.3.12. Prove that each of the following statements is true.

(a) $\mathbf{A}$ has an eigenpair $(\xi, \mathbf{x})$, where ξ is real and $\mathbf{x} > \mathbf{0}$.

(b) If λ is any eigenvalue for $\mathbf{A}$ other than ξ, then $\operatorname{Re}(\lambda) < \xi$.

(c) ξ increases when any entry in $\mathbf{A}$ is increased.

8.3.14. Let $\mathbf{A} \geq \mathbf{0}$ be an irreducible matrix, and let $a_{ij}^{(k)}$ denote entries in $\mathbf{A}^k$. Prove that $\mathbf{A}$ is primitive if and only if

$$\rho(\mathbf{A}) = \lim_{k \to \infty} \left[a_{ij}^{(k)} \right]^{1/k}.$$

8.4 STOCHASTIC MATRICES AND MARKOV CHAINS

One of the most elegant applications of the Perron–Frobenius theory is the algebraic development of the theory of finite Markov chains. The purpose of this section is to present some of the aspects of this development.

A ***stochastic matrix*** is a nonnegative matrix $\mathbf{P}_{n\times n}$ in which each row sum is equal to 1. Some authors say "*row*-stochastic" to distinguish this from the case when each column sum is 1.

A ***Markov***[92]***chain*** is a stochastic process (a set of random variables $\{X_t\}_{t=0}^{\infty}$ in which X_t has the same range $\{S_1, S_2, \ldots, S_n\}$, called the *state space*) that satisfies the *Markov property*

$$P(X_{t+1} = S_j \,|\, X_t = S_{i_t},\, X_{t-1} = S_{i_{t-1}},\, \ldots, X_0 = S_{i_0}) = P(X_{t+1} = S_j \,|\, X_t = S_{i_t})$$

for each $t = 0, 1, 2, \ldots$. Think of a Markov chain as a random chain of events that occur at discrete points $t = 0, 1, 2, \ldots$ in time, where X_t represents the state of the event that occurs at time t. For example, if a mouse moves randomly through a maze consisting of chambers $S_1, S_2, \ldots, S_n$, then X_t might represent the chamber occupied by the mouse at time t. The Markov property asserts that the process is *memoryless* in the sense that the state of the chain at the next time period depends only on the current state and not on the past history of the chain. In other words, the mouse moving through the maze obeys the Markov property if its next move doesn't depend on where in the maze it has been in the past—i.e., the mouse is not using its memory (if it has one).

To emphasize that time is considered discretely rather than continuously the phrase "*discrete-time* Markov chain" is often used, and the phrase "*finite-state* Markov chain" might be used to emphasize that the state space is finite rather than infinite.

---

[92] Andrei Andreyevich Markov (1856–1922) was born in Ryazan, Russia, and he graduated from Saint Petersburg University in 1878 where he later became a professor. Markov's early interest was number theory because this was the area of his famous teacher Pafnuty Lvovich Chebyshev (1821–1894). But when Markov discovered that he could apply his knowledge of continued fractions to probability theory, he embarked on a new course that would make him famous—enough so that there was a lunar crater named in his honor in 1964. In addition to being involved with liberal political movements (he once refused to be decorated by the Russian Czar), Markov enjoyed poetry, and in his spare time he studied poetic style. Therefore, it was no accident that led him to analyze the distribution of vowels and consonants in Pushkin's work, *Eugene Onegin,* by constructing a simple model based on the assumption that the probability that a consonant occurs at a given position in any word should depend only on whether the preceding letter is a vowel or a consonant and not on any prior history. This was the birth of the "Markov chain." Markov was wrong in one regard—he apparently believed that the only real examples of his chains were to be found in literary texts. But Markov's work in 1907 has grown to become an indispensable tool of enormous power. It launched the theory of stochastic processes that is now the foundation for understanding, explaining, and predicting phenomena in diverse areas such as atomic physics, quantum theory, biology, genetics, social behavior, economics, and finance. Markov's chains serve to underscore the point that the long-term applicability of mathematical research is impossible to predict.

Every Markov chain defines a stochastic matrix, and conversely. Let's see how this happens. The value $p_{ij}(t) = P(X_t = S_j \mid X_{t-1} = S_i)$ is the probability of being in state S_j at time t given that the chain is in state S_i at time $t-1$, so $p_{ij}(t)$ is called the ***transition probability*** of moving from S_i to S_j at time t. The matrix of transition probabilities $\mathbf{P}_{n\times n}(t) = [p_{ij}(t)]$ is clearly a nonnegative matrix, and a little thought should convince you that each row sum must be 1. Thus $\mathbf{P}(t)$ is a stochastic matrix. When the transition probabilities don't vary with time (say $p_{ij}(t) = p_{ij}$ for all t), the chain is said to be *stationary* (or *homogeneous*), and the ***transition matrix*** is the constant stochastic matrix $\mathbf{P} = [p_{ij}]$. We will make the assumption of stationarity throughout. Conversely, every stochastic matrix $\mathbf{P}_{n\times n}$ defines an n-state Markov chain because the entries p_{ij} define a set of transition probabilities, which can be interpreted as a stationary Markov chain on n states.

A ***probability distribution vector*** is defined to be a nonnegative vector $\mathbf{p}^T = (p_1, p_2, \ldots, p_n)$ such that $\sum_k p_k = 1$. (Every row in a stochastic matrix is such a vector.) For an n-state Markov chain, the kth ***step probability distribution vector*** is defined to be

$$\mathbf{p}^T(k) = \big(p_1(k), p_2(k), \ldots, p_n(k)\big), \quad k = 1, 2, \ldots, \quad \text{where} \quad p_j(k) = P(X_k = Sj).$$

In other words, $p_j(k)$ is the probability of being in the jth state after the kth step, but before the $(k+1)^{st}$ step. The ***initial distribution vector*** is

$$\mathbf{p}^T(0) = \big(p_1(0), p_2(0), \ldots, p_n(0)\big), \quad \text{where} \quad p_j(0) = P(X_0 = Sj)$$

is the probability that the chain starts in S_j.

For example, consider the Markov chain defined by placing a mouse in the 3-chamber box with connecting doors as shown in Figure 8.4.1, and suppose that the mouse moves from the chamber it occupies to another chamber by picking a door at random—say that the doors open each minute, and when they do, the mouse is forced to move by electrifying the floor of the occupied chamber.

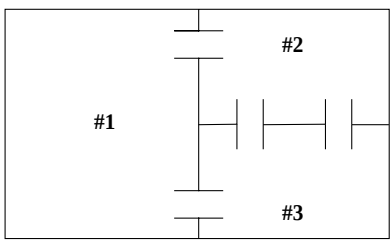


FIGURE 8.4.1

If the mouse is initially placed in chamber #2, then the initial distribution vector is $\mathbf{p}^T(0) = (0, 1, 0) = \mathbf{e}_2^T$. But if the process is started by tossing the mouse into the air so that it randomly lands in one of the chambers, then a reasonable

initial distribution is $\mathbf{p}^T(0) = (.5,\ .25,\ .25)$ because the area of chamber #1 is 50% of the box, while chambers #2 and #3 each constitute 25% of the box. The transition matrix for this Markov chain is the stochastic matrix

$$\mathbf{M} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \\ 1/3 & 2/3 & 0 \end{pmatrix}. \tag{8.4.1}$$

A standard eigenvalue calculation reveals that $\sigma(\mathbf{M}) = \{1,\ -1/3, /, -2/3\}$, so it's apparent that $\mathbf{M}$ is a nonnegative matrix having spectral radius $\rho(\mathbf{M}) = 1$. This is a feature that is shared by all stochastic matrices $\mathbf{P}_{n\times n}$ because having row sums equal to 1 means that $\|\mathbf{P}\|_\infty = 1$ or, equivalently, $\mathbf{Pe} = \mathbf{e}$, where $\mathbf{e}$ is the column of all 1's. Because $(1, \mathbf{e})$ is an eigenpair for every stochastic matrix, and because $\rho(\star) \le \|\star\|$ for every matrix norm (recall (7.1.12) on p. 497), it follows that

$$1 \le \rho(\mathbf{P}) \le \|\mathbf{P}\|_\infty = 1 \quad \Longrightarrow \quad \rho(\mathbf{P}) = 1.$$

Furthermore, $\mathbf{e}$ is a positive eigenvector associated with $\rho(\mathbf{P}) = 1$. But be careful! This doesn't mean that you necessarily can call $\mathbf{e}$ the Perron vector for $\mathbf{P}$ because $\mathbf{P}$ might not be irreducible—consider $\mathbf{P} = \begin{pmatrix} .5 & .5 \\ 0 & 1 \end{pmatrix}$.

Two important issues that arise in Markovian analysis concern the transient behavior of the chain as well as the limiting behavior. In other words, we want to accomplish the following goals.

- Describe the kth step distribution $\mathbf{p}^T(k)$ for any given initial distribution vector $\mathbf{p}^T(0)$.

- Determine whether or not $\lim_{k\to\infty} \mathbf{p}^T(k)$ exists, and if it exists, determine the value of $\lim_{k\to\infty} \mathbf{p}^T(k)$.

- If there is no limiting distribution, then determine the possibility of having a Cesàro limit

$$\lim_{k\to\infty} \left[\frac{\mathbf{p}^T(0) + \mathbf{p}^T(1) + \cdots + \mathbf{p}^T(k-1)}{k} \right].$$

  If such a limit exists, interpret its meaning, and determine its value.

The kth step distribution is easily described by using the laws of elementary probability—in particular, recall that $P(E \vee F) = P(E) + P(F)$ when E and F are mutually exclusive events, and the conditional probability of E occurring given that F occurs is $P(E \,|\, F) = P(E \wedge F)/P(F)$ (it's convenient to use $\wedge$ and $\vee$ to denote *AND* and *OR*, respectively). To determine the jth component

$p_j(1)$ in $\mathbf{p}^T(1)$ for a given $\mathbf{p}^T(0)$, write

$$\begin{aligned}
p_j(1) &= P(X_1{=}S_j) = P\Big[X_1{=}S_j \wedge (X_0{=}S_1 \vee X_0{=}S_2 \vee \cdots \vee X_0{=}S_n)\Big] \\
&= P\Big[(X_1{=}S_j \wedge X_0{=}S_1) \vee (X_1{=}S_j \wedge X_0{=}S_2) \vee \cdots \vee (X_1{=}S_j \wedge X_0{=}S_n)\Big] \\
&= \sum_{i=1}^{n} P\Big[X_1{=}S_j \wedge X_0{=}S_i\Big] = \sum_{i=1}^{n} P\Big[X_0 = S_i\Big] P\Big[X_1 = S_j \,|\, X_0 = Si\Big] \\
&= \sum_{i=1}^{n} p_i(0) p_{ij} \quad \text{for} \quad j = 1, 2, \ldots, n.
\end{aligned}$$

Consequently, $\mathbf{p}^T(1) = \mathbf{p}^T(0)\mathbf{P}$. This tells us what to expect after one step when we start with $\mathbf{p}^T(0)$. But the "no memory" Markov property tells us that the state of affairs at the end of two steps is determined by where we are at the end of the first step—it's like starting over but with $\mathbf{p}^T(1)$ as the initial distribution. In other words, it follows that $\mathbf{p}^T(2) = \mathbf{p}^T(1)\mathbf{P}$, and $\mathbf{p}^T(3) = \mathbf{p}^T(2)\mathbf{P}$, etc. Therefore, successive substitution yields

$$\mathbf{p}^T(k) = \mathbf{p}^T(k-1)\mathbf{P} = \mathbf{p}^T(k-2)\mathbf{P}^2 = \cdots = \mathbf{p}^T(0)\mathbf{P}^k,$$

and thus the kth step distribution is determined from the initial distribution and the transition matrix by the vector–matrix product

$$\mathbf{p}^T(k) = \mathbf{p}^T(0)\mathbf{P}^k. \tag{8.4.2}$$

Notice that if we adopt the notation $\mathbf{P}^k = \big[p_{ij}^{(k)}\big]$, and if we set $\mathbf{p}^T(0) = \mathbf{e}_i^T$ in (8.4.2), then we get $p_j(k) = p_{ij}^{(k)}$ for each $i = 1, 2, \ldots, n$, and thus we arrive at the following conclusion.

- The (i, j)-entry in $\mathbf{P}^k$ represents the probability of moving from S_i to S_j in exactly k steps. For this reason, $\mathbf{P}^k$ is often called the ***k-step transition matrix.***

Example 8.4.1

Let's go back to the mouse-in-the-box example, and, as suggested earlier, toss the mouse into the air so that it randomly lands somewhere in the box in Figure 8.4.1—i.e., take the initial distribution to be $\mathbf{p}^T(0) = (1/2,\ 1/4,\ 1/4)$. The transition matrix is given by (8.4.1), so the probability of finding the mouse in chamber #1 after three moves is

$$[\mathbf{p}^T(3)]_1 = [\mathbf{p}^T(0)\mathbf{M}^3]_1 = 13/54.$$

In fact, the entire third step distribution is $\mathbf{p}^T(3) = (\,13/54,\ 41/108,\ 41/108\,)$.

To analyze limiting properties of Markov chains, divide the class of stochastic matrices (and hence the class of stationary Markov chains) into four mutually exclusive categories as described below.

(1) Irreducible with $\lim_{k\to\infty}\mathbf{P}^k$ existing (i.e., $\mathbf{P}$ is primitive).
(2) Irreducible with $\lim_{k\to\infty}\mathbf{P}^k$ not existing (i.e., $\mathbf{P}$ is imprimitive).
(3) Reducible with $\lim_{k\to\infty}\mathbf{P}^k$ existing.
(4) Reducible with $\lim_{k\to\infty}\mathbf{P}^k$ not existing.

In case (1), where $\mathbf{P}$ is primitive, we know exactly what $\lim_{k\to\infty}\mathbf{P}^k$ looks like. The Perron vector for $\mathbf{P}$ is $\mathbf{e}/n$ (the uniform distribution vector), so if $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)^T$ is the Perron vector for $\mathbf{P}^T$, then

$$\lim_{k\to\infty}\mathbf{P}^k = \frac{(\mathbf{e}/n)\boldsymbol{\pi}^T}{\boldsymbol{\pi}^T(\mathbf{e}/n)} = \frac{\mathbf{e}\boldsymbol{\pi}^T}{\boldsymbol{\pi}^T\mathbf{e}} = \mathbf{e}\boldsymbol{\pi}^T = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix} > \mathbf{0} \tag{8.4.3}$$

by (8.3.10) on p. 674. Therefore, if $\mathbf{P}$ is primitive, then a limiting probability distribution exists, and it is given by

$$\lim_{k\to\infty}\mathbf{p}^T(k) = \lim_{k\to\infty}\mathbf{p}^T(0)\mathbf{P}^k = \mathbf{p}^T(0)\mathbf{e}\boldsymbol{\pi}^T = \boldsymbol{\pi}^T. \tag{8.4.4}$$

Notice that because $\sum_k p_k(0) = 1$, the term $\mathbf{p}^T(0)\mathbf{e}$ drops away, so we have the conclusion that the value of the limit is *independent* of the value of the initial distribution $\mathbf{p}^T(0)$, which isn't too surprising.

Example 8.4.2

Going back to the mouse-in-the-box example, it's easy to confirm that the transition matrix $\mathbf{M}$ in (8.4.1) is primitive, so $\lim_{k\to\infty}\mathbf{M}^k$ as well as $\lim_{k\to\infty}\mathbf{p}^T(0)$ must exist, and their values are determined by the left-hand Perron vector of $\mathbf{M}$ that can be found by calculating any nonzero vector $\mathbf{v} \in N\left(\mathbf{I} - \mathbf{M}^T\right)$ and normalizing it to produce $\boldsymbol{\pi}^T = \mathbf{v}^T/\|\mathbf{v}\|_1$. Routine computation reveals that the one solution of the homogeneous equation $(\mathbf{I}-\mathbf{M}^T)\mathbf{v} = \mathbf{0}$ is $\mathbf{v}^T = (2, 3, 3)$, so $\boldsymbol{\pi}^T = (1/8)(2, 3, 3)$, and thus

$$\lim_{k\to\infty}\mathbf{M}^k = \frac{1}{8}\begin{pmatrix} 2 & 3 & 3 \\ 2 & 3 & 3 \\ 2 & 3 & 3 \end{pmatrix} \quad \text{and} \quad \lim_{k\to\infty}\mathbf{p}^T(k) = \frac{1}{8}(2, 3, 3).$$

This limiting distribution can be interpreted as meaning that in the long run the mouse will occupy chamber #1 one-fourth of the time, while 37.5% of the time it's in chamber #2, and 37.5% of the time it's in chamber #3, and this is independent of where (or how) the process started. The mathematical justification for this statement is on p. 693.

Now consider the imprimitive case. We know that if $\mathbf{P}$ is irreducible and has $h > 1$ eigenvalues on the unit (spectral) circle, then $\lim_{k\to\infty} \mathbf{P}^k$ cannot exist (p. 674), and hence $\lim_{k\to\infty} \mathbf{p}^T(k)$ cannot exist (otherwise taking $\mathbf{p}^T(0) = \mathbf{e}_i^T$ for each i would insure that $\mathbf{P}^k$ has a limit). However, each eigenvalue on the unit circle is simple (p. 676), and this means that $\mathbf{P}$ is Cesàro summable (p. 633). Moreover, $\mathbf{e}/n$ is the Perron vector for $\mathbf{P}$, and, as pointed out in Example 8.3.2 (p. 677), if $\boldsymbol{\pi}^T = (\pi_1, \pi_2, \ldots, \pi_n)$ is the left-hand Perron vector, then

$$\lim_{k\to\infty} \frac{\mathbf{I} + \mathbf{P} + \cdots + \mathbf{P}^{k-1}}{k} = \frac{(\mathbf{e}/n)\boldsymbol{\pi}^T}{\boldsymbol{\pi}^T(\mathbf{e}/n)} = \frac{\mathbf{e}\boldsymbol{\pi}^T}{\boldsymbol{\pi}^T\mathbf{e}} = \mathbf{e}\boldsymbol{\pi}^T = \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_n \\ \pi_1 & \pi_2 & \cdots & \pi_n \\ \vdots & \vdots & & \vdots \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix},$$

which is exactly the same form as the limit (8.4.3) for the primitive case. Consequently, the kth step distributions have a Cesàro limit given by

$$\begin{aligned} \lim_{k\to\infty} \left[\frac{\mathbf{p}^T(0) + \mathbf{p}^T(1) + \cdots + \mathbf{p}^T(k-1)}{k} \right] &= \lim_{k\to\infty} \mathbf{p}^T(0) \left[\frac{\mathbf{I} + \mathbf{P} + \cdots + \mathbf{P}^{k-1}}{k} \right] \\ &= \mathbf{p}^T(0)\mathbf{e}\boldsymbol{\pi}^T = \boldsymbol{\pi}^T, \end{aligned}$$

and, just as in the primitive case (8.4.4), this Cesàro limit is independent of the initial distribution.

Let's interpret the meaning of this Cesàro limit. The analysis is essentially the same as the description outlined in the shell game in Example 7.10.8 (p. 635), but for the sake of completeness we will duplicate some of the logic here. The trick is to focus on one state, say S_j, and define a sequence of random variables $\{Z_k\}_{k=0}^{\infty}$ that count the number of visits to S_j. Let

$$Z_0 = \begin{cases} 1 & \text{if the chain starts in } S_j, \\ 0 & \text{otherwise,} \end{cases}$$

and for $i > 1$, (8.4.5)

$$Z_i = \begin{cases} 1 & \text{if the chain is in } S_j \text{ after the } i^{th} \text{ move,} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $Z_0 + Z_1 + \cdots + Z_{k-1}$ counts the number of visits to S_j before the kth move, so $(Z_0 + Z_1 + \cdots + Z_{k-1})/k$ represents the *fraction* of times that S_j is hit before the kth move. The expected (or mean) value of each Z_i is

$$E[Z_i] = 1 \cdot P(Z_i{=}1) \ + \ 0 \cdot P(Z_i{=}0) = P(Z_i{=}1) = p_j(i),$$

and, since expectation is linear, the expected fraction of times that S_j is hit before move k is

$$\begin{aligned} E\left[\frac{Z_0 + Z_1 + \cdots + Z_{k-1}}{k} \right] &= \frac{E[Z_0] + E[Z_1] + \cdots + E[Z_{k-1}]}{k} \\ &= \frac{p_j(0) + p_j(1) + \cdots + p_j(k-1)}{k} = \left[\frac{\mathbf{p}^T(0) + \mathbf{p}^T(1) + \cdots + \mathbf{p}^T(k-1)}{k} \right]_j \\ &\to \pi_j. \end{aligned}$$

In other words, the long-run fraction of time that the chain spends in S_j is π_j, which is the jth component of the Cesàro limit or, equivalently, the jth component of the left-hand Perron vector for $\mathbf{P}$.

When $\lim_{k\to\infty} \mathbf{p}^T(k)$ exists, it must be the case that

$$\lim_{k\to\infty} \mathbf{p}^T(k) = \lim_{k\to\infty} \left[\frac{\mathbf{p}^T(0)+\mathbf{p}^T(1)+\cdots+\mathbf{p}^T(k-1)}{k}\right] \quad \text{(Exercise 7.10.11, p. 639)},$$

and therefore the interpretation of the limiting distribution $\lim_{k\to\infty} \mathbf{p}^T(k)$ for the primitive case is exactly the same as the interpretation of the Cesàro limit in the imprimitive case.

Below is a summary of our findings for irreducible chains.

Irreducible Markov Chains

Let $\mathbf{P}$ be the transition probability matrix for an irreducible Markov chain on states $\{S_1, S_2, \ldots, S_n\}$ (i.e., $\mathbf{P}$ is an $n \times n$ irreducible stochastic matrix), and let $\boldsymbol{\pi}^T$ denote the left-hand Perron vector for $\mathbf{P}$. The following statements are true for every initial distribution $\mathbf{p}^T(0)$.

- The kth step transition matrix is $\mathbf{P}^k$ because the (i,j)-entry in $\mathbf{P}^k$ is the probability of moving from S_i to S_j in exactly k steps.
- The kth step distribution vector is given by $\mathbf{p}^T(k) = \mathbf{p}^T(0)\mathbf{P}^k$.
- If $\mathbf{P}$ is primitive, and if $\mathbf{e}$ denotes the column of all 1's, then

$$\lim_{k\to\infty} \mathbf{P}^k = \mathbf{e}\boldsymbol{\pi}^T \quad \text{and} \quad \lim_{k\to\infty} \mathbf{p}^T(k) = \boldsymbol{\pi}^T.$$

- If $\mathbf{P}$ is imprimitive, then

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1}}{k} = \mathbf{e}\boldsymbol{\pi}^T$$

and

$$\lim_{k\to\infty} \left[\frac{\mathbf{p}^T(0)+\mathbf{p}^T(1)+\cdots+\mathbf{p}^T(k-1)}{k}\right] = \boldsymbol{\pi}^T.$$

- Regardless of whether $\mathbf{P}$ is primitive or imprimitive, the jth component π_j of $\boldsymbol{\pi}^T$ represents the long-run fraction of time that the chain is in S_j.
- $\boldsymbol{\pi}^T$ is often called the ***stationary distribution vector*** for the chain because it is the unique distribution vector satisfying $\boldsymbol{\pi}^T\mathbf{P} = \boldsymbol{\pi}^T$.

Example 8.4.3

Periodic Chains. Consider an electronic switch that can be in one of three states $\{S_1, S_2, S_3\}$, and suppose that the switch changes states on regular clock cycles. If the switch is in either S_1 or S_3, then it must change to S_2 on the next clock cycle, but if the switch is in S_2, then there is an equal likelihood that it changes to S_1 or S_3 on the next clock cycle. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix},$$

and it's not difficult to see that $\mathbf{P}$ is irreducible (because $\mathcal{G}(\mathbf{P})$ is strongly connected) and imprimitive (because $\sigma(\mathbf{P}) = \{\pm 1, 0\}$). Since the left-hand Perron vector is $\boldsymbol{\pi}^T = (.25, .5, .25)$, the long-run expectation is that the switch should be in S_1 25% of the time, in S_2 50% of the time, and in S_3 25% of the time, and this agrees with what common sense tells us. Furthermore, notice that the switch cannot be in just any position at any given clock cycle because if the chain starts in either S_1 or S_3, then it must be in S_2 on every odd-numbered cycle, and it can occupy S_1 or S_3 only on even-numbered cycles. The situation is similar, but with reversed parity, when the chain starts in S_2. In other words, the chain is periodic in the sense that the states can be occupied only at periodic points in time. In this example the period of the chain is 2, and this is the same as the index of imprimitivity. This is no accident. The Frobenius form for imprimitive matrices on p. 680 can be used to prove that this is true in general. Consequently, an irreducible Markov chain is said to be a ***periodic chain*** when its transition matrix $\mathbf{P}$ is imprimitive (with the period of the chain being the index of imprimitivity for $\mathbf{P}$), and an irreducible Markov chain for which $\mathbf{P}$ is primitive is called an ***aperiodic chain.*** The shell game in Example 7.10.8 (p. 635) is a periodic Markov chain that is similar to the one in this example.

Because the Perron–Frobenius theorem is not directly applicable to reducible chains (chains for which $\mathbf{P}$ is a reducible matrix), the strategy for analyzing reducible chains is to deflate the situation, as much as possible, back to the irreducible case as described below.

If $\mathbf{P}$ is reducible, then, by definition, there exists a permutation matrix $\mathbf{Q}$ and square matrices $\mathbf{X}$ and $\mathbf{Z}$ such that

$$\mathbf{Q}^T\mathbf{P}\mathbf{Q} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}. \text{ For convenience, denote this by writing } \mathbf{P} \sim \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}.$$

If $\mathbf{X}$ or $\mathbf{Z}$ is reducible, then another symmetric permutation can be performed to produce

$$\begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \sim \begin{pmatrix} \mathbf{R} & \mathbf{S} & \mathbf{T} \\ \mathbf{0} & \mathbf{U} & \mathbf{V} \\ \mathbf{0} & \mathbf{0} & \mathbf{W} \end{pmatrix}, \quad \text{where } \mathbf{R}, \ \mathbf{U}, \text{ and } \mathbf{W} \text{ are square.}$$

Repeating this process eventually yields

$$\mathbf{P} \sim \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1k} \\ \mathbf{0} & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2k} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_{kk} \end{pmatrix}, \quad \text{where each } \mathbf{X}_{ii} \text{ is irreducible or } \mathbf{X}_{ii} = [0]_{1\times 1}.$$

Finally, if there exist rows having nonzero entries *only* in diagonal blocks, then symmetrically permute all such rows to the bottom to produce

$$\mathbf{P} \sim \left(\begin{array}{cccc|cccc} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{rr} & \mathbf{P}_{1,r+1} & \mathbf{P}_{1,r+2} & \cdots & \mathbf{P}_{1m} \\ \mathbf{0} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2r} & \mathbf{P}_{2,r+1} & \mathbf{P}_{2,r+2} & \cdots & \mathbf{P}_{2m} \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{rr} & \mathbf{P}_{r,r+1} & \mathbf{P}_{r,r+2} & \cdots & \mathbf{P}_{rm} \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{P}_{r+1,r+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{P}_{r+2,r+2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{P}_{mm} \end{array}\right), \tag{8.4.6}$$

where each $\mathbf{P}_{11}, \ldots, \mathbf{P}_{rr}$ is either irreducible or $[0]_{1\times 1}$, and $\mathbf{P}_{r+1,r+1}, \ldots, \mathbf{P}_{mm}$ are irreducible (they can't be zero because each has row sums equal to 1). As mentioned on p. 671, the effect of a symmetric permutation is simply to relabel nodes in $\mathcal{G}(\mathbf{P})$ or, equivalently, to reorder the states in the chain. When the states of a chain have been reordered so that $\mathbf{P}$ assumes the form on the right-hand side of (8.4.6), we say that $\mathbf{P}$ is in the ***canonical form for reducible matrices.*** When $\mathbf{P}$ is in canonical form, the subset of states corresponding to $\mathbf{P}_{kk}$ for $1 \leq k \leq r$ is called the kth ***transient class*** (because once left, a transient class can't be reentered), and the subset of states corresponding to $\mathbf{P}_{r+j,r+j}$ for $j \geq 1$ is called the jth ***ergodic class.*** Each ergodic class is an irreducible Markov chain unto itself that is imbedded in the larger reducible chain. From now on, we will assume that the states in our reducible chains have been ordered so that $\mathbf{P}$ is in canonical form.

The results on p. 676 guarantee that if an irreducible stochastic matrix $\mathbf{P}$ has h eigenvalues on the unit circle, then these h eigenvalues are the hth roots of unity, and each is a simple eigenvalue for $\mathbf{P}$. The same can't be said for reducible stochastic matrices, but the canonical form (8.4.6) allows us to prove the next best thing as discussed below.

Unit Eigenvalues

The ***unit eigenvalues*** for a stochastic matrix are defined to be those eigenvalues that are on the unit circle. For every stochastic matrix $\mathbf{P}_{n\times n}$, the following statements are true.

- Every unit eigenvalue of $\mathbf{P}$ is semisimple.
- Every unit eigenvalue has form $\lambda = \mathrm{e}^{2k\pi\mathrm{i}/h}$ for some $k < h \leq n$.
- In particular, $\rho(\mathbf{P}) = 1$ is always a semisimple eigenvalue of $\mathbf{P}$.

Proof. If $\mathbf{P}$ is irreducible, then there is nothing to prove because, as proved on p. 676, the unit eigenvalues are roots of unity, and each unit eigenvalue is simple. If $\mathbf{P}$ is reducible, suppose that a symmetric permutation has been performed so that $\mathbf{P}$ is in the canonical form (8.4.6), and observe that

$$\rho(\mathbf{P}_{kk}) < 1 \quad \text{for each } k = 1, 2, \ldots, r. \tag{8.4.7}$$

This is certainly true when $\mathbf{P}_{kk} = [0]_{1\times 1}$, so suppose that $\mathbf{P}_{kk}$ $(1 \leq k \leq r)$ is irreducible. Because there must be blocks $\mathbf{P}_{kj}$, $j \neq k$, that have nonzero entries, it follows that

$$\mathbf{P}_{kk}\mathbf{e} \leq \mathbf{e} \quad \text{and} \quad \mathbf{P}_{kk}\mathbf{e} \neq \mathbf{e}, \quad \text{where } \mathbf{e} \text{ is the column of all 1's.}$$

If $\rho(\mathbf{P}_{kk}) = 1$, then the observation in Example 8.3.1 (p. 674) forces $\mathbf{P}_{kk}\mathbf{e} = \mathbf{e}$, which is impossible, and thus $\rho(\mathbf{P}_{kk}) < 1$. Consequently, the unit eigenvalues for $\mathbf{P}$ are the collection of the unit eigenvalues of the irreducible matrices $\mathbf{P}_{r+1,r+1}, \ldots, \mathbf{P}_{mm}$. But each unit eigenvalue of $\mathbf{P}_{r+i,r+i}$ is simple and is a root of unity. Consequently, if λ is a unit eigenvalue for $\mathbf{P}$, then it must be some root of unity, and although it might be repeated because it appears in the spectrum of more than one $\mathbf{P}_{r+i,r+i}$, it must nevertheless be the case that $\mathit{alg\ mult}_{\mathbf{P}}(\lambda) = \mathit{geo\ mult}_{\mathbf{P}}(\lambda)$, so λ is a semisimple eigenvalue of $\mathbf{P}$. ∎

We know from the discussion on p. 633 that a matrix $\mathbf{A} \in \mathcal{C}^{n\times n}$ is Cesàro summable if and only if $\rho(\mathbf{A}) < 1$ or $\rho(\mathbf{A}) = 1$ with each eigenvalue on the unit circle being semisimple. We just proved that the latter holds for all stochastic matrices $\mathbf{P}$, so we have in fact established the following powerful statement concerning all stochastic matrices.

All Stochastic Matrices Are Summable

Every stochastic matrix $\mathbf{P}$ is Cesàro summable. That is,

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1}}{k} \quad \text{exists for all stochastic matrices } \mathbf{P},$$

and, as discussed on p. 633, the value of the limit is the (spectral) projector $\mathbf{G}$ onto $N(\mathbf{I}-\mathbf{P})$ along $R(\mathbf{I}-\mathbf{P})$.

Since we already know the structure and interpretation of the Cesàro limit when $\mathbf{P}$ is an irreducible stochastic matrix (p. 693), all that remains in order to complete the picture is to analyze the nature of $\lim_{k\to\infty}(\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1})/k$ for the reducible case.

Suppose that $\mathbf{P} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}$ is a reducible stochastic matrix that is in the canonical form (8.4.6), where

$$\mathbf{T}_{11} = \begin{pmatrix} \mathbf{P}_{11} & \cdots & \mathbf{P}_{rr} \\ & \ddots & \vdots \\ & & \mathbf{P}_{rr} \end{pmatrix}, \qquad \mathbf{T}_{12} = \begin{pmatrix} \mathbf{P}_{1,r+1} & \cdots & \mathbf{P}_{1m} \\ \vdots & & \vdots \\ \mathbf{P}_{r,r+1} & \cdots & \mathbf{P}_{rm} \end{pmatrix},$$

and (8.4.8)

$$\mathbf{T}_{22} = \begin{pmatrix} \mathbf{P}_{r+1,r+1} & & \\ & \ddots & \\ & & \mathbf{P}_{mm} \end{pmatrix}.$$

We know from (8.4.7) that $\rho(\mathbf{P}_{kk}) < 1$ for each $k = 1, 2, \ldots, r$, so it follows that $\rho(\mathbf{T}_{11}) < 1$, and hence

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{T}_{11}+\cdots+\mathbf{T}_{11}^{k-1}}{k} = \lim_{k\to\infty} \mathbf{T}_{11}^{k} = \mathbf{0} \quad \text{(recall Exercise 7.10.11 on p. 639).}$$

Furthermore, $\mathbf{P}_{r+1,r+1}, \ldots, \mathbf{P}_{mm}$ are each irreducible stochastic matrices, so if $\boldsymbol{\pi}_j^T$ is the left-hand Perron vector for $\mathbf{P}_{jj}$, $r+1 \le j \le m$, then our previous results (p. 693) tell us that

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{T}_{22}+\cdots+\mathbf{T}_{22}^{k-1}}{k} = \begin{pmatrix} \mathbf{e}\boldsymbol{\pi}_{r+1}^T & & \\ & \ddots & \\ & & \mathbf{e}\boldsymbol{\pi}_m^T \end{pmatrix} = \mathbf{E}. \tag{8.4.9}$$

Furthermore, it's clear from the results on p. 674 that $\lim_{k\to\infty} \mathbf{T}_{22}^k$ exists if and only if $\mathbf{P}_{r+1,r+1}, \ldots, \mathbf{P}_{mm}$ are each primitive, in which case $\lim_{k\to\infty} \mathbf{T}_{22}^k = \mathbf{E}$.

Therefore, the limits, be they Cesàro or ordinary (if it exists), all have the form

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1}}{k} = \begin{pmatrix} \mathbf{0} & \mathbf{Z} \\ \mathbf{0} & \mathbf{E} \end{pmatrix} = \mathbf{G} = \lim_{k\to\infty} \mathbf{P}^k \text{ (when it exists).}$$

To determine the precise nature of $\mathbf{Z}$, use the fact that $R(\mathbf{G}) = N(\mathbf{I}-\mathbf{P})$ (because $\mathbf{G}$ is the projector onto $N(\mathbf{I}-\mathbf{P})$ along $R(\mathbf{I}-\mathbf{P})$) to write

$$(\mathbf{I}-\mathbf{P})\mathbf{G} = \mathbf{0} \implies \begin{pmatrix} \mathbf{I}-\mathbf{T}_{11} & -\mathbf{T}_{12} \\ \mathbf{0} & \mathbf{I}-\mathbf{T}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{Z} \\ \mathbf{0} & \mathbf{E} \end{pmatrix} = \mathbf{0} \implies (\mathbf{I}-\mathbf{T}_{11})\mathbf{Z} = \mathbf{T}_{12}\mathbf{E}.$$

Since $\mathbf{I}-\mathbf{T}_{11}$ is nonsingular (because $\rho(\mathbf{T}_{11}) < 1$ by (8.4.7)), it follows that

$$\mathbf{Z} = (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}\mathbf{E},$$

and thus the following results concerning limits of reducible chains are produced.

Reducible Markov Chains

If the states in a reducible Markov chain have been ordered to make the transition matrix assume the canonical form

$$\mathbf{P} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{T}_{22} \end{pmatrix}$$

that is described in (8.4.6) and (8.4.8), and if $\boldsymbol{\pi}_j^T$ is the left-hand Perron vector for $\mathbf{P}_{jj}$ $(r+1 \le j \le m)$, then $\mathbf{I}-\mathbf{T}_{11}$ is nonsingular, and

$$\lim_{k\to\infty} \frac{\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1}}{k} = \begin{pmatrix} \mathbf{0} & (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}\mathbf{E} \\ \mathbf{0} & \mathbf{E} \end{pmatrix},$$

where

$$\mathbf{E} = \begin{pmatrix} \mathbf{e}\boldsymbol{\pi}_{r+1}^T & & \\ & \ddots & \\ & & \mathbf{e}\boldsymbol{\pi}_m^T \end{pmatrix}.$$

Furthermore, $\lim_{k\to\infty} \mathbf{P}^k$ exists if and only if the stochastic matrices $\mathbf{P}_{r+1,r+1}, \ldots, \mathbf{P}_{mm}$ in (8.4.6) are each primitive, in which case

$$\lim_{k\to\infty} \mathbf{P}^k = \begin{pmatrix} \mathbf{0} & (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}\mathbf{E} \\ \mathbf{0} & \mathbf{E} \end{pmatrix}. \tag{8.4.10}$$

The preceding analysis shows that every reducible chain eventually gets *absorbed* (trapped) into one of the ergodic classes—i.e., into a subchain defined by $\mathbf{P}_{r+j,r+j}$ for some $j \geq 1$. If $\mathbf{P}_{r+j,r+j}$ is primitive, then the chain settles down to a steady-state defined by the left-hand Perron vector of $\mathbf{P}_{r+j,r+j}$, but if $\mathbf{P}_{r+j,r+j}$ is imprimitive, then the process will oscillate in the jth ergodic class forever. There is not much more that can be said about the limit, but there are still important questions concerning which ergodic class the chain will end up in and how long it takes to get there. This time the answer depends on where the chain starts—i.e., on the initial distribution.

For convenience, let $\mathcal{T}_i$ denote the ith transient class, and let $\mathcal{E}_j$ be the jth ergodic class. Suppose that the chain starts in a particular transient state—say we start in the pth state of $\mathcal{T}_i$. Since the question at hand concerns only which ergodic class is hit but not what happens after it's entered, we might as well convert every state in each ergodic class into a trap by setting $\mathbf{P}_{r+j,r+j} = \mathbf{I}$ for each $j \geq 1$ in (8.4.6). The transition matrix for this modified chain is $\widetilde{\mathbf{P}} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, and it follows from (8.4.10) that $\lim_{k\to\infty} \widetilde{\mathbf{P}}^k$ exists and has the form

$$\lim_{k\to\infty} \widetilde{\mathbf{P}}^k = \begin{pmatrix} \mathbf{0} & (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \left(\begin{array}{cccc|cccc} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L}_{1,1} & \mathbf{L}_{1,2} & \cdots & \mathbf{L}_{1s} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L}_{2,1} & \mathbf{L}_{2,2} & \cdots & \mathbf{L}_{2s} \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L}_{r,1} & \mathbf{L}_{r,2} & \cdots & \mathbf{L}_{rs} \\ \hline \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{array}\right).$$

Consequently, the (p,q)-entry in block $\mathbf{L}_{ij}$ represents the probability of eventually hitting the qth state in $\mathcal{E}_j$ given that we start from the pth state in $\mathcal{T}_i$. Therefore, if $\mathbf{e}$ is the vector of all 1's, then the probability of eventually entering somewhere in $\mathcal{E}_j$ is given by

- $P(\text{absorption into } \mathcal{E}_j \,|\, \text{start in } p^{th} \text{ state of } \mathcal{T}_i) = \sum_k \big[\mathbf{L}_{ij}\big]_{pk} = \big[\mathbf{L}_{ij}\mathbf{e}\big]_p$.

If $\mathbf{p}_i^T(0)$ is an initial distribution for starting in the various states of $\mathcal{T}_i$, then

- $P\big(\text{absorption into } \mathcal{E}_j \,|\, \mathbf{p}_i^T(0)\big) = \mathbf{p}_i^T(0)\mathbf{L}_{ij}\mathbf{e}$.

To determine the expected number of steps required to first hit an ergodic state, proceed as follows. Count the number of times the chain is in transient state S_j given that it starts in transient state S_i by reapplying the argument given in (8.4.5) on p. 692. That is, given that the chain starts in S_i, let

$$Z_0 = \begin{cases} 1 & \text{if } S_i = S_j, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad Z_k = \begin{cases} 1 & \text{if the chain is in } S_j \text{ after step } k, \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$E[Z_k] = 1 \cdot P(Z_k{=}1) \;+\; 0 \cdot P(Z_k{=}0) = P(Z_k{=}1) = \left[\mathbf{T}_{11}^k\right]_{ij},$$

and since $\sum_{k=0}^{\infty} Z_k$ is the total number of times the chain is in S_j, we have

$$\begin{aligned} E[\#\text{ times in } S_j|\text{ start in } S_i] &= E\left[\sum_{k=0}^{\infty} Z_k\right] = \sum_{k=0}^{\infty} E\left[Z_k\right] = \sum_{k=0}^{\infty}\left[\mathbf{T}_{11}^k\right]_{ij} \\ &= \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\right]_{ij} \quad (\text{because } \rho\left(\mathbf{T}_{11}\right) < 1). \end{aligned}$$

Summing this over all transient states produces the expected number of times the chain is in *some* transient state, which is the same as the expected number of times before first hitting an ergodic state. In other words,

- $E[\#\text{ steps until absorption}\,|\,\text{start in } i^{th} \text{ transient state}] = \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e}\right]_i.$

Example 8.4.4

Absorbing Markov Chains. It's often the case in practical applications that there is only one transient class, and the ergodic classes are just single *absorbing states* (states such that once they are entered, they are never left). If the single transient class contains r states, and if there are s absorbing states, then the canonical form for the transition matrix is

$$\mathbf{P} = \left(\begin{array}{ccc|ccc} p_{11} & \cdots & p_{1r} & p_{1,r+1} & \cdots & p_{1s} \\ \vdots & & \vdots & \vdots & & \vdots \\ p_{r1} & \cdots & p_{rr} & p_{r,r+1} & \cdots & p_{rs} \\ \hline 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{array}\right) \quad \text{and} \quad \mathbf{L}_{ij} = \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}\right]_{ij}.$$

The preceding analysis specializes to say that every absorbing chain must eventually reach one of its absorbing states. The probability of being absorbed into the jth absorbing state (which is state S_{r+j}) given that the chain starts in the ith transient state (which is S_i) is

$$P(\text{absorption into } S_{r+j}|\,\text{start in } S_i \text{ for } 1 \le i \le r) = \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}\right]_{ij},$$

while the expected time until absorption is

$$E[\#\text{ steps until absorption}\,|\,\text{start in } S_i] = \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e}\right]_i,$$

and the amount of time spent in S_j is

$$E[\#\text{ times in } S_j|\text{ start in } S_i] = \left[(\mathbf{I}-\mathbf{T}_{11})^{-1}\right]_{ij}.$$

Example 8.4.5

Fail-Safe System. Consider a system that has two independent controls, A and B, that can prevent the system from being destroyed. The system is activated at discrete points in time $t_1, t_2, t_3, \ldots,$ and the system is considered to be "under control" if either control A or B holds at the time of activation. The system is destroyed if A and B fail simultaneously.

▷ For example, an automobile has two independent braking systems—one is operated by a foot pedal, whereas the "emergency brake" is operated by a hand lever. The automobile is "under control" if at least one braking system is operative when you try to stop, but a crash occurs if both braking systems fail simultaneously.

If one of the controls fails at some activation point but the other control holds, then the defective control is repaired before the next activation. If a control holds at time $t = t_k$, then it is considered to be 90% reliable at $t = t_{k+1}$, but if a control fails at time $t = t_k$, then its untested replacement is considered to be only 60% reliable at $t = t_{k+1}$.

Problem: Can the system be expected to run indefinitely without every being destroyed? If not, how long is the system expected to run before destruction occurs?

Solution: This is a four-state Markov chain with the states being the controls that hold at any particular time of activation. In other words the state space is the set of pairs (a, b) in which

$$a = \begin{cases} 1 & \text{if A holds,} \\ 0 & \text{if A fails,} \end{cases} \qquad \text{and} \qquad b = \begin{cases} 1 & \text{if B holds,} \\ 0 & \text{if B fails.} \end{cases}$$

State $(0, 0)$ is absorbing, and the transition matrix (in canonical form) is

$$\mathbf{P} = \begin{array}{c} \\ (1,1) \\ (1,0) \\ (0,1) \\ (0,0) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,0) & (0,1) & (0,0) \end{array} \\ \begin{pmatrix} .81 & .09 & .09 & .01 \\ .54 & .36 & .06 & .04 \\ .54 & .06 & .36 & .04 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

with

$$\mathbf{T}_{11} = \begin{pmatrix} .81 & .09 & .09 \\ .54 & .36 & .06 \\ .54 & .06 & .36 \end{pmatrix} \qquad \text{and} \qquad \mathbf{T}_{12} = \begin{pmatrix} .01 \\ .04 \\ .04 \end{pmatrix}.$$

The fact that $\lim_{k\to\infty} \mathbf{P}^k$ exists and is given by

$$\lim_{k\to\infty} \mathbf{P}^k = \begin{pmatrix} \mathbf{0} & (\mathbf{I} - \mathbf{T}_{11})^{-1}\mathbf{T}_{12} \\ \mathbf{0} & 1 \end{pmatrix}$$

makes it clear that the absorbing state must eventually be reached. In other words, this proves the validity of the popular belief that "if something can go wrong, then it eventually will." Rounding to three significant figures produces

$$(\mathbf{I}-\mathbf{T}_{11})^{-1} = \begin{pmatrix} 44.6 & 6.92 & 6.92 \\ 41.5 & 8.02 & 6.59 \\ 41.5 & 6.59 & 8.02 \end{pmatrix} \quad \text{and} \quad (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e} = \begin{pmatrix} 58.4 \\ 56.1 \\ 56.1 \end{pmatrix},$$

so the mean time to failure starting with two proven controls is slightly more than 58 steps, while the mean time to failure starting with one untested control and one proven control is just over 56 steps. The difference here doesn't seem significant, but consider what happens when only one control is used in the system. In this case, there are only two states in the chain, 1 (meaning that the control holds) and 0 (meaning that it doesn't). The transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 0 \end{array} \\ \begin{pmatrix} .9 & .1 \\ 0 & 1 \end{pmatrix} \end{array},$$

so now the mean time to failure is only $(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e} = 10$ steps. It's interesting to consider what happens when three independent control are used. How much more security does your intuition tell you that you should have? See Exercise 8.4.8.

Exercises for section 8.4

8.4.1. Find the stationary distribution for $\mathbf{P} = \begin{pmatrix} 1/4 & 0 & 0 & 3/4 \\ 3/8 & 1/4 & 3/8 & 0 \\ 1/3 & 1/6 & 1/6 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$. Does this stationary distribution represent a limiting distribution in the regular sense or only in the Cesàro sense?

8.4.2. A ***doubly-stochastic matrix*** is a nonnegative matrix $\mathbf{P}_{n\times n}$ having all row sums as well as all column sums equal to 1. For an irreducible n-state Markov chain whose transition matrix is doubly stochastic, what is the long-run proportion of time spent in each state? What form do $\lim_{k\to\infty}(\mathbf{I}+\mathbf{P}+\cdots+\mathbf{P}^{k-1})/k$ and $\lim_{k\to\infty}\mathbf{P}^k$ (if it exists) have? **Note:** The purpose of this exercise is to show that doubly-stochastic matrices are not very interesting from a Markov-chain point of view. However, there is an interesting theoretical result (due to G. Birkhoff in 1946) that says the set of $n\times n$ doubly-stochastic matrices forms a convex polyhedron in $\Re^{n\times n}$ with the permutation matrices as the vertices.

8.4.3. Explain why $rank\,(\mathbf{I}-\mathbf{P}) = n-1$ for every irreducible stochastic matrix $\mathbf{P}_{nn}$. Give an example to show that this need not be the case for reducible stochastic matrices.

8.4.4. Prove that the left-hand Perron vector for an irreducible stochastic matrix $\mathbf{P}_{n\times n}$ $(n>1)$ is given by

$$\boldsymbol{\pi}^T = \frac{1}{\sum_{i=1}^{n} P_i}\,(P_1,\, P_2,\, \ldots,\, P_n),$$

where P_i is the ith principal minor determinant of order $n-1$ in $\mathbf{I}-\mathbf{P}$. **Hint:** What is $[\mathrm{adj}\,(\mathbf{A})]\mathbf{A}$ if $\mathbf{A}$ is singular?

8.4.5. Let $\mathbf{P}_{n\times n}$ be an irreducible stochastic matrix, and let $\mathbf{Q}_{k\times k}$ be a principal submatrix of $\mathbf{I}-\mathbf{P}$, where $1 \le k < n$. Prove that $\rho\,(\mathbf{Q}) < 1$.

8.4.6. Let $\mathbf{P}_{n\times n}$ be an irreducible stochastic matrix, and let $\mathbf{Q}_{k\times k}$ be a principal submatrix of $\mathbf{I}-\mathbf{P}$, where $1 \le k < n$. Explain why $\mathbf{Q}$ is an M-matrix as defined and discussed on p. 626.

8.4.7. Let $\mathbf{P}_{n\times n}$ $(n>1)$ be an irreducible stochastic matrix. Explain why all principal minors of order $1 \le k < n$ in $\mathbf{I}-\mathbf{P}$ are positive.

8.4.8. Use the same assumptions that are used for the fail-safe system described in Example 8.4.5, but use three controls, A, B, and C, instead of two. Determine the mean time to failure starting with three proven controls, two proven but one untested control, and three untested controls.

8.4.9. A mouse is placed in one chamber of the box shown in Figure 8.4.1 on p. 688, and a cat is placed in another chamber. Each minute the doors to the chambers are opened just long enough to allow movement from one chamber to an adjacent chamber. Half of the time when the doors are opened, the cat doesn't leave the chamber it occupies. The same is true for the mouse. When either the cat or mouse moves, a door is chosen at random to pass through.

(a) Explain why the cat and mouse must eventually end up in the same chamber, and determine the expected number of steps for this to occur.

(b) Determine the probability that the cat will catch the mouse in chamber $\#j$ for each $j = 1, 2, 3$.

Solutions for Chapter 1

Solutions for exercises in section 1. 2

1.2.1. $(1, 0, 0)$

1.2.2. $(1, 2, 3)$

1.2.3. $(1, 0, -1)$

1.2.4. $(-1/2, 1/2, 0, 1)$

1.2.5. $\left(\begin{array}{c|c|c} 2 & -4 & 3 \\ 4 & -7 & 4 \\ 5 & -8 & 4 \end{array}\right)$

1.2.6. Every row operation is reversible. In particular the "inverse" of any row operation is again a row operation of the same type.

1.2.7. $\frac{\pi}{2}, \pi, 0$

1.2.8. The third equation in the triangularized form is $0x_3 = 1$, which is impossible to solve.

1.2.9. The third equation in the triangularized form is $0x_3 = 0$, and all numbers are solutions. This means that you can start the back substitution with any value whatsoever and consequently produce infinitely many solutions for the system.

1.2.10. $\alpha = -3$, $\beta = \frac{11}{2}$, and $\gamma = -\frac{3}{2}$

1.2.11. (a) If $x_i =$ the number initially in chamber $\#i$, then

$$\begin{aligned} .4x_1 + 0x_2 + 0x_3 + .2x_4 &= 12 \\ 0x_1 + .4x_2 + .3x_3 + .2x_4 &= 25 \\ 0x_1 + .3x_2 + .4x_3 + .2x_4 &= 26 \\ .6x_1 + .3x_2 + .3x_3 + .4x_4 &= 37 \end{aligned}$$

and the solution is $x_1 = 10$, $x_2 = 20$, $x_3 = 30$, and $x_4 = 40$.
(b) 16, 22, 22, 40

1.2.12. To interchange rows i and j, perform the following sequence of Type II and Type III operations.

$$\begin{array}{ll} R_j \leftarrow R_j + R_i & \text{(replace row } j \text{ by the sum of row } j \text{ and } i) \\ R_i \leftarrow R_i - R_j & \text{(replace row } i \text{ by the difference of row } i \text{ and } j) \\ R_j \leftarrow R_j + R_i & \text{(replace row } j \text{ by the sum of row } j \text{ and } i) \\ R_i \leftarrow -R_i & \text{(replace row } i \text{ by its negative)} \end{array}$$

1.2.13. (a) This has the effect of interchanging the order of the unknowns—x_j and x_k are permuted. (b) The solution to the new system is the same as the

solution to the old system except that the solution for the jth unknown of the new system is $\hat{x}_j = \frac{1}{\alpha} x_j$. This has the effect of "changing the units" of the jth unknown. (c) The solution to the new system is the same as the solution for the old system except that the solution for the kth unknown in the new system is $\hat{x}_k = x_k - \alpha x_j$.

1.2.14. $h_{ij} = \frac{1}{i+j-1}$

1.2.16. If $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$ are two different solutions, then

$$\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2} = \begin{pmatrix} \frac{x_1+y_1}{2} \\ \frac{x_2+y_2}{2} \\ \vdots \\ \frac{x_m+y_m}{2} \end{pmatrix}$$

is a third solution different from both $\mathbf{x}$ and $\mathbf{y}$.

Solutions for exercises in section 1. 3

1.3.1. $(1, 0, -1)$

1.3.2. $(2, -1, 0, 0)$

1.3.3. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$

Solutions for exercises in section 1. 4

1.4.2. Use $y'(t_k) = y'_k \approx \dfrac{y_{k+1} - y_{k-1}}{2h}$ and $y''(t_k) = y''_k \approx \dfrac{y_{k-1} - 2y_k + y_{k+1}}{h^2}$ to write

$$f(t_k) = f_k = y''_k - y'_k \approx \frac{2y_{k-1} - 4y_k + 2y_{k+1}}{2h^2} - \frac{hy_{k+1} - hy_{k-1}}{2h^2}, \quad k = 1, 2, \ldots, n,$$

with $y_0 = y_{n+1} = 0$. These discrete approximations form the tridiagonal system

$$\begin{pmatrix} -4 & 2-h & & & \\ 2+h & -4 & 2-h & & \\ & \ddots & \ddots & \ddots & \\ & & 2+h & -4 & 2-h \\ & & & 2+h & -4 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = 2h^2 \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix}.$$

Solutions for exercises in section 1. 5

1.5.1. (a) $(0,-1)$ (c) $(1,-1)$ (e) $\left(\frac{1}{1.001}, \frac{-1}{1.001}\right)$

1.5.2. (a) $(0,1)$ (b) $(2,1)$ (c) $(2,1)$ (d) $\left(\frac{2}{1.0001}, \frac{1.0003}{1.0001}\right)$

1.5.3. Without PP: (1.01, 1.03) With PP: (1, 1) Exact: (1, 1)

1.5.4. (a)
$$\left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ .500 & .333 & .250 & .333 \\ .333 & .250 & .200 & .200 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .083 & .083 & .166 \\ 0 & .083 & .089 & .089 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .083 & .083 & .166 \\ 0 & 0 & .006 & -.077 \end{array}\right) \quad z = -.077/.006 = -12.8,$$

$$y = (.166 - .083z)/.083 = 14.8, \quad x = .333 - (.5y + .333z) = -2.81$$

(b)
$$\left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ .500 & .333 & .250 & .333 \\ .333 & .250 & .200 & .200 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 1 & .666 & .500 & .666 \\ 1 & .751 & .601 & .601 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .166 & .167 & .333 \\ 0 & .251 & .268 & .268 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .251 & .268 & .268 \\ 0 & .166 & .167 & .333 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .251 & .268 & .268 \\ 0 & 0 & -.01 & .156 \end{array}\right) \quad z = -.156/.01 = -15.6,$$

$$y = (.268 - .268z)/.251 = 17.7, \quad x = .333 - (.5y + .333z) = -3.33$$

(c)
$$\left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ .500 & .333 & .250 & .333 \\ .333 & .250 & .200 & .200 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 1 & .666 & .500 & .666 \\ 1 & .751 & .601 & .601 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .166 & .167 & .333 \\ 0 & .251 & .268 & .268 \end{array}\right) \longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .994 & 1 & 1.99 \\ 0 & .937 & 1 & 1 \end{array}\right)$$

$$\longrightarrow \left(\begin{array}{ccc|c} 1 & .500 & .333 & .333 \\ 0 & .994 & 1 & 1.99 \\ 0 & 0 & .057 & -.880 \end{array}\right) \quad z = -.88/.057 = -15.4,$$

$$y = (1.99 - z)/.994 = 17.5, \quad x = .333 - (.5y + .333z) = -3.29$$

(d) $x = -3, \quad y = 16, \quad z = -14$

1.5.5. (a)

$$\begin{aligned} .0055x + .095y + 960z &= 5000 \\ .0011x + .\ 01y + 112z &= 600 \\ .0093x + .025y + 560z &= 3000 \end{aligned}$$

(b) 3-digit solution $= (55,900$ lbs. silica, $8,600$ lbs. iron, 4.04 lbs. gold). Exact solution (to 10 digits) $= (56,753.68899,\ 8,626.560726,\ 4.029511918)$. The relative error (rounded to 3 digits) is $e_r = 1.49 \times 10^{-2}$.

(c) Let $u = x/2000$, $v = y/1000$, and $w = 12z$ to obtain the system

$$\begin{aligned} 11u + 95v + \quad 80w &= 5000 \\ 2.2u + 10v + 9.33w &= 600 \\ 18.6u + 25v + 46.7w &= 3000. \end{aligned}$$

(d) 3-digit solution $=$ (28.5 tons silica, 8.85 half-tons iron, 48.1 troy oz. gold). Exact solution (to 10 digits) $=$ (28.82648317, 8.859282804, 48.01596023). The relative error (rounded to 3 digits) is $e_r = 5.95 \times 10^{-3}$. So, partial pivoting applied to the column-scaled system yields higher relative accuracy than partial pivoting applied to the unscaled system.

1.5.6. (a) $(-8.1, -6.09) =$ 3-digit solution with partial pivoting but no scaling.
(b) No! Scaled partial pivoting produces the exact solution—the same as with complete pivoting.

1.5.7. (a) 2^{n-1} (b) 2

(c) This is a famous example that shows that there are indeed cases where partial pivoting will fail due to the large growth of some elements during elimination, but complete pivoting will be successful because all elements remain relatively small and of the same order of magnitude.

1.5.8. Use the fact that with partial pivoting no multiplier can exceed 1 together with the triangle inequality $|\alpha + \beta| \le |\alpha| + |\beta|$ and proceed inductively.

Solutions for exercises in section 1. 6

1.6.1. (a) There are no 5-digit solutions. (b) This doesn't help—there are now infinitely many 5-digit solutions. (c) 6-digit solution $= (1.23964, -1.3)$ and exact solution $= (1, -1)$ (d) $r_1 = r_2 = 0$ (e) $r_1 = -10^{-6}$ and $r_2 = 10^{-7}$ (f) Even if computed residuals are 0, you can't be sure you have the exact solution.

1.6.2. (a) $(1, -1.0015)$ (b) Ill-conditioning guarantees that the solution will be very sensitive to *some* small perturbation but not necessarily to *every* small perturbation. It is usually difficult to determine beforehand those perturbations for which an ill-conditioned system will not be sensitive, so one is forced to be pessimistic whenever ill-conditioning is suspected.

1.6.3. (a) $m_1(5) = m_2(5) = -1.2519$, $m_1(6) = -1.25187$, and $m_2(6) = -1.25188$
(c) An optimally well-conditioned system represents orthogonal (i.e., perpendicular) lines, planes, etc.

1.6.4. They rank as (b) = Almost optimally well-conditioned. (a) = Moderately well-conditioned. (c) = Badly ill-conditioned.

1.6.5. Original solution $= (1, 1, 1)$. Perturbed solution $= (-238, 490, -266)$. System is ill-conditioned.

Solutions for Chapter 2

Solutions for exercises in section 2. 1

2.1.1. (a) $\begin{pmatrix} 1 & 2 & 3 & 3 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}$ is one possible answer. Rank $= 3$ and the basic columns are $\{\mathbf{A}_{*1}, \mathbf{A}_{*2}, \mathbf{A}_{*4}\}$. (b) $\begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 2 \\ 0 & 0 & -8 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ is one possible answer. Rank $= 3$ and every column in $\mathbf{A}$ is basic.

(c) $\begin{pmatrix} 2 & 1 & 1 & 3 & 0 & 4 & 1 \\ 0 & 0 & 2 & -2 & 1 & -3 & 3 \\ 0 & 0 & 0 & 0 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ is one possible answer. The rank is 3, and the basic columns are $\{\mathbf{A}_{*1}, \mathbf{A}_{*3}, \mathbf{A}_{*5}\}$.

2.1.2. (c) and (d) are in row echelon form.

2.1.3. (a) Since any row or column can contain at most one pivot, the number of pivots cannot exceed the number of rows nor the number of columns. (b) A zero row cannot contain a pivot. (c) If one row is a multiple of another, then one of them can be annihilated by the other to produce a zero row. Now the result of the previous part applies. (d) One row can be annihilated by the associated combination of row operations. (e) If a column is zero, then there are fewer than n basic columns because each basic column must contain a pivot.

2.1.4. (a) $rank\,(\mathbf{A}) = 3$ (b) 3-digit $rank\,(\mathbf{A}) = 2$ (c) With PP, 3-digit $rank\,(\mathbf{A}) = 3$

2.1.5. 15

2.1.6. (a) No, consider the form $\left(\begin{array}{ccc|c} * & * & * & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & * \end{array}\right)$ (b) Yes—in fact, $\mathbf{E}$ is a row echelon form obtainable from $\mathbf{A}$.

Solutions for exercises in section 2. 2

2.2.1. (a) $\begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and $\mathbf{A}_{*3} = 2\mathbf{A}_{*1} + \frac{1}{2}\mathbf{A}_{*2}$

(b) $\begin{pmatrix} 1 & \frac{1}{2} & 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ and

$\mathbf{A}_{*2} = \frac{1}{2}\mathbf{A}_{*1}, \quad \mathbf{A}_{*4} = 2\mathbf{A}_{*1} - \mathbf{A}_{*3}, \quad \mathbf{A}_{*6} = 2\mathbf{A}_{*1} - 3\mathbf{A}_{*5}, \quad \mathbf{A}_{*7} = \mathbf{A}_{*3} + \mathbf{A}_{*5}$

2.2.2. No.

2.2.3. The same would have to hold in $\mathbf{E_A}$, and there you can see that this means not all columns can be basic. Remember, $rank\,(\mathbf{A}) =$ number of basic columns.

2.2.4. (a) $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$ $\mathbf{A}_{*3}$ is almost a combination of $\mathbf{A}_{*1}$ and $\mathbf{A}_{*2}$. In particular, $\mathbf{A}_{*3} \approx -\mathbf{A}_{*1} + 2\mathbf{A}_{*2}$.

2.2.5. $\mathbf{E}_{*1} = 2\mathbf{E}_{*2} - \mathbf{E}_{*3}$ and $\mathbf{E}_{*2} = \frac{1}{2}\mathbf{E}_{*1} + \frac{1}{2}\mathbf{E}_{*3}$

Solutions for exercises in section 2. 3

2.3.1. (a), (b)—There is no need to do any arithmetic for this one because the right-hand side is entirely zero so that you know (0,0,0) is automatically one solution. (d), (f)

2.3.3. It is always true that $rank\,(\mathbf{A}) \leq rank[\mathbf{A}|\mathbf{b}] \leq m$. Since $rank\,(\mathbf{A}) = m$, it follows that $rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A})$.

2.3.4. Yes—Consistency implies that $\mathbf{b}$ and $\mathbf{c}$ are each combinations of the basic columns in $\mathbf{A}$. If $\mathbf{b} = \sum \beta_i \mathbf{A}_{*b_i}$ and $\mathbf{c} = \sum \gamma_i \mathbf{A}_{*b_i}$ where the $\mathbf{A}_{*b_i}$'s are the basic columns, then $\mathbf{b} + \mathbf{c} = \sum(\beta_i + \gamma_i)\mathbf{A}_{*b_i} = \sum \xi_i \mathbf{A}_{*b_i}$, where $\xi_i = \beta_i + \gamma_i$ so that $\mathbf{b} + \mathbf{c}$ is also a combination of the basic columns in $\mathbf{A}$.

2.3.5. Yes—because the 4×3 system $\alpha + \beta x_i + \gamma x_i^2 = y_i$ obtained by using the four given points (x_i, y_i) is consistent.

2.3.6. The system is inconsistent using 5-digits but consistent when 6-digits are used.

2.3.7. If x, y, and z denote the number of pounds of the respective brands applied, then the following constraints must be met.

$$\begin{aligned} \text{total \# units of phosphorous} &= 2x + y + z = 10 \\ \text{total \# units of potassium} &= 3x + 3y = 9 \\ \text{total \# units of nitrogen} &= 5x + 4y + z = 19 \end{aligned}$$

Since this is a consistent system, the recommendation can be satisfied exactly. Of course, the solution tells how much of each brand to apply.

2.3.8. No—if one or more such rows were ever present, how could you possibly eliminate all of them with row operations? You could eliminate all but one, but then there is no way to eliminate the last remaining one, and hence it would have to appear in the final form.

Solutions for exercises in section 2. 4

2.4.1. (a) $x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}$ (b) $y \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{pmatrix}$ (c) $x_3 \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$

(d) The trivial solution is the only solution.

2.4.2. $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -\frac{1}{2} \\ 0 \end{pmatrix}$

2.4.3. $x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -2 \\ 0 \\ -1 \\ 1 \\ 0 \end{pmatrix}$

2.4.4. $rank\,(\mathbf{A}) = 3$

2.4.5. (a) 2—because the maximum rank is 4. (b) 5—because the minimum rank is 1.

2.4.6. Because $r = rank\,(\mathbf{A}) \leq m < n \implies n - r > 0$.

2.4.7. There are many different correct answers. One approach is to answer the question "What must $\mathbf{E_A}$ look like?" The form of the general solution tells you that $rank\,(\mathbf{A}) = 2$ and that the first and third columns are basic. Consequently, $\mathbf{E_A} = \begin{pmatrix} 1 & \alpha & 0 & \beta \\ 0 & 0 & 1 & \gamma \\ 0 & 0 & 0 & 0 \end{pmatrix}$ so that $x_1 = -\alpha x_2 - \beta x_4$ and $x_3 = -\gamma x_4$ gives rise to the general solution $x_2 \begin{pmatrix} -\alpha \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -\beta \\ 0 \\ -\gamma \\ 1 \end{pmatrix}$. Therefore, $\alpha = 2$, $\beta = 3$, and $\gamma = -2$. Any matrix $\mathbf{A}$ obtained by performing row operations to $\mathbf{E_A}$ will be the coefficient matrix for a homogeneous system with the desired general solution.

2.4.8. If $\sum_i x_{f_i}\mathbf{h}_i$ is the general solution, then there must exist scalars α_i and β_i such that $\mathbf{c}_1 = \sum_i \alpha_i \mathbf{h}_i$ and $\mathbf{c}_2 = \sum_i \beta_i \mathbf{h}_i$. Therefore, $\mathbf{c}_1 + \mathbf{c}_2 = \sum_i (\alpha_i + \beta_i)\mathbf{h}_i$, and this shows that $\mathbf{c}_1 + \mathbf{c}_2$ is the solution obtained when the free variables x_{f_i} assume the values $x_{f_i} = \alpha_i + \beta_i$.

Solutions for exercises in section 2. 5

2.5.1. (a) $\begin{pmatrix} 1 \\ 0 \\ 2 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 0 \\ -1 \\ 1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + y \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \end{pmatrix}$

(c) $\begin{pmatrix} 2 \\ -1 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}$ (d) $\begin{pmatrix} 3 \\ -3 \\ -1 \end{pmatrix}$

2.5.2. From Example 2.5.1, the solutions of the linear equations are:

$$\begin{aligned} x_1 &= 1 - x_3 - 2x_4 \\ x_2 &= 1 - x_3 \\ x_3 &\text{ is free} \\ x_4 &\text{ is free} \\ x_5 &= -1 \end{aligned}$$

Substitute these into the two constraints to get $x_3 = \pm 1$ and $x_4 = \pm 1$. Thus there are exactly four solutions:

$$\left\{ \begin{pmatrix} -2 \\ 0 \\ 1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 4 \\ 2 \\ -1 \\ -1 \\ -1 \end{pmatrix} \right\}$$

2.5.3. (a) $\{(3,0,4),\ (2,1,5),\ (1,2,6),\ (0,3,7)\}$ See the solution to Exercise 2.3.7 for the underlying system. (b) $(3,0,4)$ costs \$15 and is least expensive.

2.5.4. (a) Consistent for all α. (b) $\alpha \neq 3$, in which case the solution is $(1,-1,0)$.

(c) $\alpha = 3$, in which case the general solution is $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ -\frac{3}{2} \\ 1 \end{pmatrix}$.

2.5.5. No

2.5.6.

$$\mathbf{E_A} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}_{m \times n}$$

2.5.7. See the solution to Exercise 2.4.7.

2.5.8. (a) $\begin{pmatrix} -.3976 \\ 0 \\ 1 \end{pmatrix} + y \begin{pmatrix} -.7988 \\ 1 \\ 0 \end{pmatrix}$ (b) There are no solutions in this case.

(c) $\begin{pmatrix} 1.43964 \\ -2.3 \\ 1 \end{pmatrix}$

Solutions for exercises in section 2. 6

2.6.1. (a) $(1/575)(383, 533, 261, 644, -150, -111)$
2.6.2. $(1/211)(179, 452, 36)$
2.6.3. (18, 10)
2.6.4. (a) 4 (b) 6 (c) 7 loops but only 3 simple loops. (d) Show that $rank\,([\mathbf{A}|\mathbf{b}]) = 3$ (g) 5/6

I fear explanations explanatory of things explained.
— *Abraham Lincoln* (1809–1865)

Solutions for Chapter 3

Solutions for exercises in section 3. 2

3.2.1. (a) $\mathbf{X} = \begin{pmatrix} 0 & 1 \\ 2 & 3 \end{pmatrix}$ (b) $x = -\frac{1}{2}$, $y = -6$, and $z = 0$

3.2.2. (a) Neither (b) Skew symmetric (c) Symmetric (d) Neither

3.2.3. The 3×3 zero matrix trivially satisfies all conditions, and it is the only possible answer for part (a). The only possible answers for (b) are real symmetric matrices. There are many nontrivial possibilities for (c).

3.2.4. $\mathbf{A} = \mathbf{A}^T$ and $\mathbf{B} = \mathbf{B}^T \implies (\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T = \mathbf{A} + \mathbf{B}$. Yes—the skew-symmetric matrices are also closed under matrix addition.

3.2.5. (a) $\mathbf{A} = -\mathbf{A}^T \implies a_{ij} = -a_{ji}$. If $i = j$, then $a_{jj} = -a_{jj} \implies a_{jj} = 0$.

(b) $\mathbf{A} = -\mathbf{A}^* \implies a_{ij} = -\overline{a_{ji}}$. If $i = j$, then $a_{jj} = -\overline{a_{jj}}$. Write $a_{jj} = x + \mathrm{i}y$ to see that $a_{jj} = -\overline{a_{jj}} \implies x + \mathrm{i}y = -x + \mathrm{i}y \implies x = 0 \implies a_{jj}$ is pure imaginary.

(c) $\mathbf{B}^* = (\mathrm{i}\mathbf{A})^* = -\mathrm{i}\mathbf{A}^* = -\mathrm{i}\overline{\mathbf{A}}^T = -\mathrm{i}\mathbf{A}^T = -\mathrm{i}\mathbf{A} = -\mathbf{B}$.

3.2.6. (a) Let $\mathbf{S} = \mathbf{A} + \mathbf{A}^T$ and $\mathbf{K} = \mathbf{A} - \mathbf{A}^T$. Then $\mathbf{S}^T = \mathbf{A}^T + \mathbf{A}^{T^T} = \mathbf{A}^T + \mathbf{A} = \mathbf{S}$. Likewise, $\mathbf{K}^T = \mathbf{A}^T - \mathbf{A}^{T^T} = \mathbf{A}^T - \mathbf{A} = -\mathbf{K}$.

(b) $\mathbf{A} = \frac{\mathbf{S}}{2} + \frac{\mathbf{K}}{2}$ is one such decomposition. To see it is unique, suppose $\mathbf{A} = \mathbf{X} + \mathbf{Y}$, where $\mathbf{X} = \mathbf{X}^T$ and $\mathbf{Y} = -\mathbf{Y}^T$. Thus, $\mathbf{A}^T = \mathbf{X}^T + \mathbf{Y}^T = \mathbf{X} - \mathbf{Y} \implies \mathbf{A} + \mathbf{A}^T = 2\mathbf{X}$, so that $\mathbf{X} = \frac{\mathbf{A}+\mathbf{A}^T}{2} = \frac{\mathbf{S}}{2}$. A similar argument shows that $\mathbf{Y} = \frac{\mathbf{A}-\mathbf{A}^T}{2} = \frac{\mathbf{K}}{2}$.

3.2.7. (a) $[(\mathbf{A}+\mathbf{B})^*]_{ij} = [\overline{\mathbf{A}+\mathbf{B}}]_{ji} = [\overline{\mathbf{A}} + \overline{\mathbf{B}}]_{ji} = [\overline{\mathbf{A}}]_{ji} + [\overline{\mathbf{B}}]_{ji} = [\mathbf{A}^*]_{ij} + [\mathbf{B}^*]_{ij} = [\mathbf{A}^* + \mathbf{B}^*]_{ij}$

(b) $[(\alpha\mathbf{A})^*]_{ij} = [\overline{\alpha\mathbf{A}}]_{ji} = [\bar{\alpha}\overline{\mathbf{A}}]_{ji} = \bar{\alpha}[\overline{\mathbf{A}}]_{ji} = \bar{\alpha}[\mathbf{A}^*]_{ij}$

3.2.8. $k \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$

Solutions for exercises in section 3. 3

3.3.1. Functions (b) and (f) are linear. For example, to check if (b) is linear, let $\mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, and check if $f(\mathbf{A}+\mathbf{B}) = f(\mathbf{A}) + f(\mathbf{B})$ and

$f(\alpha\mathbf{A}) = \alpha f(\mathbf{A})$. Do so by writing

$$f(\mathbf{A}+\mathbf{B}) = f\begin{pmatrix} a_1+b_1 \\ a_2+b_2 \end{pmatrix} = \begin{pmatrix} a_2+b_2 \\ a_1+b_1 \end{pmatrix} = \begin{pmatrix} a_2 \\ a_1 \end{pmatrix} + \begin{pmatrix} b_2 \\ b_1 \end{pmatrix} = f(\mathbf{A}) + f(\mathbf{B}),$$

$$f(\alpha\mathbf{A}) = f\begin{pmatrix} \alpha a_1 \\ \alpha a_2 \end{pmatrix} = \begin{pmatrix} \alpha a_2 \\ \alpha a_1 \end{pmatrix} = \alpha\begin{pmatrix} a_2 \\ a_1 \end{pmatrix} = \alpha f(\mathbf{A}).$$

3.3.2. Write $f(\mathbf{x}) = \sum_{i=1}^n \xi_i x_i$. For all points $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, and for all scalars α, it is true that

$$\begin{aligned} f(\alpha\mathbf{x}+\mathbf{y}) &= \sum_{i=1}^n \xi_i(\alpha x_i + y_i) = \sum_{i=1}^n \xi_i \alpha x_i + \sum_{i=1}^n \xi_i y_i \\ &= \alpha\sum_{i=1}^n \xi_i x_i + \sum_{i=1}^n \xi_i y_i = \alpha f(\mathbf{x}) + f(\mathbf{y}). \end{aligned}$$

3.3.3. There are many possibilities. Two of the simplest and most common are Hooke's law for springs that says that $F = kx$ (see Example 3.2.1) and Newton's second law that says that $F = ma$ (i.e., force = mass × acceleration).

3.3.4. They are all linear. To see that rotation is linear, use trigonometry to deduce that if $\mathbf{p} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, then $f(\mathbf{p}) = \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, where

$$\begin{aligned} u_1 &= (\cos\theta)x_1 - (\sin\theta)x_2 \\ u_2 &= (\sin\theta)x_1 + (\cos\theta)x_2. \end{aligned}$$

f is linear because this is a special case of Example 3.3.2. To see that reflection is linear, write $\mathbf{p} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $f(\mathbf{p}) = \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix}$. Verification of linearity is straightforward. For the projection function, use the Pythagorean theorem to conclude that if $\mathbf{p} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, then $f(\mathbf{p}) = \frac{x_1+x_2}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Linearity is now easily verified.

Solutions for exercises in section 3. 4

3.4.1. Refer to the solution for Exercise 3.3.4. If $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{P}$ denote the matrices associated with the rotation, reflection, and projection, respectively, then

$$\mathbf{Q} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad \mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

3.4.2. Refer to the solution for Exercise 3.4.1 and write

$$\mathbf{RQ} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ -\sin\theta & -\cos\theta \end{pmatrix}.$$

If $Q(\mathbf{x})$ is the rotation function and $R(\mathbf{x})$ is the reflection function, then the composition is

$$R\Big(Q(\mathbf{x})\Big) = \begin{pmatrix} (\cos\theta)x_1 - (\sin\theta)x_2 \\ -(\sin\theta)x_1 - (\cos\theta)x_2 \end{pmatrix}.$$

3.4.3. Refer to the solution for Exercise 3.4.1 and write

$$\begin{aligned} \mathbf{PQR} &= \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \cos\theta + \sin\theta & \sin\theta - \cos\theta \\ \cos\theta + \sin\theta & \sin\theta - \cos\theta \end{pmatrix}. \end{aligned}$$

Therefore, the composition of the three functions in the order asked for is

$$P\Big(Q\Big(R(\mathbf{x})\Big)\Big) = \frac{1}{2} \begin{pmatrix} (\cos\theta + \sin\theta)x_1 + (\sin\theta - \cos\theta)x_2 \\ (\cos\theta + \sin\theta)x_1 + (\sin\theta - \cos\theta)x_2 \end{pmatrix}.$$

Solutions for exercises in section 3. 5

3.5.1. (a) $\mathbf{AB} = \begin{pmatrix} 10 & 15 \\ 12 & 8 \\ 28 & 52 \end{pmatrix}$ (b) $\mathbf{BA}$ does not exist (c) $\mathbf{CB}$ does not exist

(d) $\mathbf{C}^T\mathbf{B} = \begin{pmatrix} 10 & 31 \end{pmatrix}$ (e) $\mathbf{A}^2 = \begin{pmatrix} 13 & -1 & 19 \\ 16 & 13 & 12 \\ 36 & -17 & 64 \end{pmatrix}$ (f) $\mathbf{B}^2$ does not exist

(g) $\mathbf{C}^T\mathbf{C} = 14$ (h) $\mathbf{CC}^T = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}$ (i) $\mathbf{BB}^T = \begin{pmatrix} 5 & 8 & 17 \\ 8 & 16 & 28 \\ 17 & 28 & 58 \end{pmatrix}$

(j) $\mathbf{B}^T\mathbf{B} = \begin{pmatrix} 10 & 23 \\ 23 & 69 \end{pmatrix}$ (k) $\mathbf{C}^T\mathbf{AC} = 76$

3.5.2. (a) $\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 0 & 2 \\ 2 & 2 & 0 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} 3 \\ 10 \\ -2 \end{pmatrix}$ (b) $\mathbf{s} = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$

(c) $\mathbf{b} = \mathbf{A}_{*1} - 2\mathbf{A}_{*2} + 3\mathbf{A}_{*3}$

3.5.3. (a) $\mathbf{EA} = \begin{pmatrix} \mathbf{A}_{1*} \\ \mathbf{A}_{2*} \\ 3\mathbf{A}_{1*} + \mathbf{A}_{3*} \end{pmatrix}$ (b) $\mathbf{AE} = (\, \mathbf{A}_{*1} + 3\mathbf{A}_{*3} \quad \mathbf{A}_{*2} \quad \mathbf{A}_{*3} \,)$

3.5.4. (a) $\mathbf{A}_{*j}$ (b) $\mathbf{A}_{i*}$ (c) a_{ij}

3.5.5. $\mathbf{Ax} = \mathbf{Bx} \;\forall\; \mathbf{x} \implies \mathbf{Ae}_j = \mathbf{Be}_j \;\forall\; \mathbf{e}_j \implies \mathbf{A}_{*j} = \mathbf{B}_{*j} \;\forall\; j \implies \mathbf{A} = \mathbf{B}$.
(The symbol $\forall$ is mathematical shorthand for the phrase "for all.")

3.5.6. The limit is the zero matrix.

3.5.7. If $\mathbf{A}$ is $m \times p$ and $\mathbf{B}$ is $p \times n$, write the product as

$$\mathbf{AB} = (\, \mathbf{A}_{*1} \quad \mathbf{A}_{*2} \quad \cdots \quad \mathbf{A}_{*p} \,) \begin{pmatrix} \mathbf{B}_{1*} \\ \mathbf{B}_{2*} \\ \vdots \\ \mathbf{B}_{p*} \end{pmatrix} = \mathbf{A}_{*1}\mathbf{B}_{1*} + \mathbf{A}_{*2}\mathbf{B}_{2*} + \cdots + \mathbf{A}_{*p}\mathbf{B}_{p*}$$

$$= \sum_{k=1}^{p} \mathbf{A}_{*k}\mathbf{B}_{k*}.$$

3.5.8. (a) $[\mathbf{AB}]_{ij} = \mathbf{A}_{i*}\mathbf{B}_{*j} = (\, 0 \quad \cdots \quad 0 \quad a_{ii} \quad \cdots \quad a_{in} \,) \begin{pmatrix} b_{1j} \\ \vdots \\ b_{jj} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ is 0 when $i > j$.

(b) When $i = j$, the only nonzero term in the product $\mathbf{A}_{i*}\mathbf{B}_{*i}$ is $a_{ii}b_{ii}$.
(c) Yes.

3.5.9. Use $[\mathbf{AB}]_{ij} = \sum_k a_{ik}b_{kj}$ along with the rules of differentiation to write

$$\frac{d[\mathbf{AB}]_{ij}}{dt} = \frac{d\left(\sum_k a_{ik}b_{kj}\right)}{dt} = \sum_k \frac{d(a_{ik}b_{kj})}{dt}$$

$$= \sum_k \left(\frac{da_{ik}}{dt} b_{kj} + a_{ik}\frac{db_{kj}}{dt} \right) = \sum_k \frac{da_{ik}}{dt} b_{kj} + \sum_k a_{ik}\frac{db_{kj}}{dt}$$

$$= \left[\frac{d\mathbf{A}}{dt}\mathbf{B} \right]_{ij} + \left[\mathbf{A}\frac{d\mathbf{B}}{dt} \right]_{ij} = \left[\frac{d\mathbf{A}}{dt}\mathbf{B} + \mathbf{A}\frac{d\mathbf{B}}{dt} \right]_{ij}.$$

3.5.10. (a) $[\mathbf{Ce}]_i =$ the total number of paths *leaving* node i.
(b) $[\mathbf{e}^T\mathbf{C}]_i =$ the total number of paths *entering* node i.

3.5.11. At time t, the concentration of salt in tank i is $\frac{x_i(t)}{V}$ lbs/gal. For tank 1,

$$\begin{aligned}\frac{dx_1}{dt} &= \frac{\text{lbs}}{\text{sec}} \text{ coming in} - \frac{\text{lbs}}{\text{sec}} \text{ going out} = 0\frac{\text{lbs}}{\text{sec}} - \left(r\frac{\text{gal}}{\text{sec}} \times \frac{x_1(t)}{V}\frac{\text{lbs}}{\text{gal}}\right)\\ &= -\frac{r}{V}x_1(t)\frac{\text{lbs}}{\text{sec}}.\end{aligned}$$

For tank 2,

$$\begin{aligned}\frac{dx_2}{dt} &= \frac{\text{lbs}}{\text{sec}} \text{ coming in} - \frac{\text{lbs}}{\text{sec}} \text{ going out} = \frac{r}{V}x_1(t)\frac{\text{lbs}}{\text{sec}} - \left(r\frac{\text{gal}}{\text{sec}} \times \frac{x_2(t)}{V}\frac{\text{lbs}}{\text{gal}}\right)\\ &= \frac{r}{V}x_1(t)\frac{\text{lbs}}{\text{sec}} - \frac{r}{V}x_2(t)\frac{\text{lbs}}{\text{sec}} = \frac{r}{V}\Big(x_1(t) - x_2(t)\Big),\end{aligned}$$

and for tank 3,

$$\begin{aligned}\frac{dx_3}{dt} &= \frac{\text{lbs}}{\text{sec}} \text{ coming in} - \frac{\text{lbs}}{\text{sec}} \text{ going out} = \frac{r}{V}x_2(t)\frac{\text{lbs}}{\text{sec}} - \left(r\frac{\text{gal}}{\text{sec}} \times \frac{x_3(t)}{V}\frac{\text{lbs}}{\text{gal}}\right)\\ &= \frac{r}{V}x_2(t)\frac{\text{lbs}}{\text{sec}} - \frac{r}{V}x_3(t)\frac{\text{lbs}}{\text{sec}} = \frac{r}{V}\Big(x_2(t) - x_3(t)\Big).\end{aligned}$$

This is a system of three linear first-order differential equations

$$\begin{aligned}\frac{dx_1}{dt} &= \tfrac{r}{V}\Big(-x_1(t) \qquad\qquad\qquad \Big)\\ \frac{dx_2}{dt} &= \tfrac{r}{V}\Big(\quad x_1(t) - x_2(t) \qquad \Big)\\ \frac{dx_3}{dt} &= \tfrac{r}{V}\Big(\qquad\qquad x_2(t) - x_3(t)\Big)\end{aligned}$$

that can be written as a single matrix differential equation

$$\begin{pmatrix} dx_1/dt \\ dx_2/dt \\ dx_3/dt \end{pmatrix} = \frac{r}{V}\begin{pmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}.$$

Solutions for exercises in section 3. 6

3.6.1.

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_1 + \mathbf{A}_{12}\mathbf{B}_2 + \mathbf{A}_{13}\mathbf{B}_3 \\ \mathbf{A}_{21}\mathbf{B}_1 + \mathbf{A}_{22}\mathbf{B}_2 + \mathbf{A}_{23}\mathbf{B}_3 \end{pmatrix}$$

$$= \left(\begin{array}{cc} -10 & -19 \\ -10 & -19 \\ \hline -1 & -1 \end{array}\right)$$

3.6.2. Use block multiplication to verify $\mathbf{L}^2 = \mathbf{I}$—be careful not to commute any of the terms when forming the various products.

3.6.3. Partition the matrix as $\mathbf{A} = \begin{pmatrix} \mathbf{I} & \mathbf{C} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$, where $\mathbf{C} = \frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ and observe that $\mathbf{C}^2 = \mathbf{C}$. Use this together with block multiplication to conclude that

$$\mathbf{A}^k = \begin{pmatrix} \mathbf{I} & \mathbf{C} + \mathbf{C}^2 + \mathbf{C}^3 + \cdots + \mathbf{C}^k \\ \mathbf{0} & \mathbf{C}^k \end{pmatrix} = \begin{pmatrix} \mathbf{I} & k\mathbf{C} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}.$$

Therefore, $\mathbf{A}^{300} = \begin{pmatrix} 1 & 0 & 0 & 100 & 100 & 100 \\ 0 & 1 & 0 & 100 & 100 & 100 \\ 0 & 0 & 1 & 100 & 100 & 100 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}$.

3.6.4. $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*\mathbf{A}^{**} = \mathbf{A}^*\mathbf{A}$ and $(\mathbf{AA}^*)^* = \mathbf{A}^{**}\mathbf{A}^* = \mathbf{AA}^*$.

3.6.5. $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T = \mathbf{BA} = \mathbf{AB}$. It is easy to construct a 2×2 example to show that this need not be true when $\mathbf{AB} \neq \mathbf{BA}$.

3.6.6.

$$\begin{aligned} [(\mathbf{D}+\mathbf{E})\mathbf{F}]_{ij} &= (\mathbf{D}+\mathbf{E})_{i*}\mathbf{F}_{*j} = \sum_k [\mathbf{D}+\mathbf{E}]_{ik}[\mathbf{F}]_{kj} = \sum_k \left([\mathbf{D}]_{ik} + [\mathbf{E}]_{ik}\right)[\mathbf{F}]_{kj} \\ &= \sum_k \left([\mathbf{D}]_{ik}[\mathbf{F}]_{kj} + [\mathbf{E}]_{ik}[\mathbf{F}]_{kj}\right) = \sum_k [\mathbf{D}]_{ik}[\mathbf{F}]_{kj} + \sum_k [\mathbf{E}]_{ik}[\mathbf{F}]_{kj} \\ &= \mathbf{D}_{i*}\mathbf{F}_{*j} + \mathbf{E}_{i*}\mathbf{F}_{*j} = [\mathbf{DF}]_{ij} + [\mathbf{EF}]_{ij} \\ &= [\mathbf{DF}+\mathbf{EF}]_{ij}. \end{aligned}$$

3.6.7. If a matrix $\mathbf{X}$ did indeed exist, then

$$\begin{aligned} \mathbf{I} = \mathbf{AX} - \mathbf{XA} \quad &\Longrightarrow \quad trace\,(\mathbf{I}) = trace\,(\mathbf{AX} - \mathbf{XA}) \\ &\Longrightarrow \quad n = trace\,(\mathbf{AX}) - trace\,(\mathbf{XA}) = 0, \end{aligned}$$

which is impossible.

3.6.8. (a) $\mathbf{y}^T\mathbf{A} = \mathbf{b}^T \implies (\mathbf{y}^T\mathbf{A})^T = \mathbf{b}^{T^T} \implies \mathbf{A}^T\mathbf{y} = \mathbf{b}$. This is an $n \times m$ system of equations whose coefficient matrix is $\mathbf{A}^T$. (b) They are the same.

3.6.9. Draw a transition diagram similar to that in Example 3.6.3 with North and South replaced by ON and OFF, respectively. Let x_k be the proportion of switches in the ON state, and let y_k be the proportion of switches in the OFF state after k clock cycles have elapsed. According to the given information,

$$x_k = x_{k-1}(.1) + y_{k-1}(.3)$$
$$y_k = x_{k-1}(.9) + y_{k-1}(.7)$$

so that $\mathbf{p}_k = \mathbf{p}_{k-1}\mathbf{P}$, where

$$\mathbf{p}_k = (\,x_k \quad y_k\,) \quad \text{and} \quad \mathbf{P} = \begin{pmatrix} .1 & .9 \\ .3 & .7 \end{pmatrix}.$$

Just as in Example 3.6.3, $\mathbf{p}_k = \mathbf{p}_0\mathbf{P}^k$. Compute a few powers of $\mathbf{P}$ to find

$$\mathbf{P}^2 = \begin{pmatrix} .280 & .720 \\ .240 & .760 \end{pmatrix}, \quad \mathbf{P}^3 = \begin{pmatrix} .244 & .756 \\ .252 & .748 \end{pmatrix}$$
$$\mathbf{P}^4 = \begin{pmatrix} .251 & .749 \\ .250 & .750 \end{pmatrix}, \quad \mathbf{P}^5 = \begin{pmatrix} .250 & .750 \\ .250 & .750 \end{pmatrix}$$

and deduce that $\mathbf{P}^\infty = \lim_{k\to\infty}\mathbf{P}^k = \begin{pmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{pmatrix}$. Thus

$$\mathbf{p}_k \to \mathbf{p}_0\mathbf{P}^\infty = (\,\tfrac{1}{4}(x_0+y_0) \quad \tfrac{3}{4}(x_0+y_0)\,) = (\,\tfrac{1}{4} \quad \tfrac{3}{4}\,).$$

For practical purposes, the device can be considered to be in equilibrium after about 5 clock cycles—regardless of the initial proportions.

3.6.10. $(\,-4 \quad 1 \quad -6 \quad 5\,)$

3.6.11. (a) $trace\,(\mathbf{ABC}) = trace\,(\mathbf{A}\{\mathbf{BC}\}) = trace\,(\{\mathbf{BC}\}\mathbf{A}) = trace\,(\mathbf{BCA})$. The other equality is similar. (b) Use almost any set of 2×2 matrices to construct an example that shows equality need not hold. (c) Use the fact that $trace\left(\mathbf{C}^T\right) = trace\,(\mathbf{C})$ for all square matrices to conclude that

$$\begin{aligned} trace\left(\mathbf{A}^T\mathbf{B}\right) &= trace\left((\mathbf{A}^T\mathbf{B})^T\right) = trace\left(\mathbf{B}^T\mathbf{A}^{T^T}\right) \\ &= trace\left(\mathbf{B}^T\mathbf{A}\right) = trace\left(\mathbf{AB}^T\right). \end{aligned}$$

3.6.12. (a) $\mathbf{x}^T\mathbf{x} = 0 \iff \sum_{k=1}^n x_i^2 = 0 \iff x_i = 0$ for each $i \iff \mathbf{x} = \mathbf{0}$.

(b)
$$\begin{aligned} trace\left(\mathbf{A}^T\mathbf{A}\right) = 0 &\iff \sum_i [\mathbf{A}^T\mathbf{A}]_{ii} = 0 \iff \sum_i (\mathbf{A}^T)_{i*}\mathbf{A}_{*i} = 0 \\ &\iff \sum_i\sum_k [\mathbf{A}^T]_{ik}[\mathbf{A}]_{ki} = 0 \iff \sum_i\sum_k [\mathbf{A}]_{ki}[\mathbf{A}]_{ki} = 0 \\ &\iff \sum_i\sum_k [\mathbf{A}]_{ki}^2 = 0 \\ &\iff [\mathbf{A}]_{ki} = 0 \text{ for each } k \text{ and } i \iff \mathbf{A} = \mathbf{0} \end{aligned}$$

Solutions for exercises in section 3. 7

3.7.1. (a) $\begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix}$ (b) Singular (c) $\begin{pmatrix} 2 & -4 & 3 \\ 4 & -7 & 4 \\ 5 & -8 & 4 \end{pmatrix}$ (d) Singular

(e) $\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$

3.7.2. Write the equation as $(\mathbf{I}-\mathbf{A})\mathbf{X}=\mathbf{B}$ and compute

$$\mathbf{X}=(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}=\begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \end{pmatrix}=\begin{pmatrix} 2 & 4 \\ -1 & -2 \\ 3 & 3 \end{pmatrix}.$$

3.7.3. In each case, the given information implies that $rank\,(\mathbf{A})<n$—see the solution for Exercise 2.1.3.

3.7.4. (a) If $\mathbf{D}$ is diagonal, then $\mathbf{D}^{-1}$ exists if and only if each $d_{ii}\neq 0$, in which case

$$\begin{pmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{pmatrix}^{-1}=\begin{pmatrix} 1/d_{11} & 0 & \cdots & 0 \\ 0 & 1/d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/d_{nn} \end{pmatrix}.$$

(b) If $\mathbf{T}$ is triangular, then $\mathbf{T}^{-1}$ exists if and only if each $t_{ii}\neq 0$. If $\mathbf{T}$ is upper (lower) triangular, then $\mathbf{T}^{-1}$ is also upper (lower) triangular with $[\mathbf{T}^{-1}]_{ii}=1/t_{ii}$.

3.7.5. $\left(\mathbf{A}^{-1}\right)^T=\left(\mathbf{A}^T\right)^{-1}=\mathbf{A}^{-1}$.

3.7.6. Start with $\mathbf{A}(\mathbf{I}-\mathbf{A})=(\mathbf{I}-\mathbf{A})\mathbf{A}$ and apply $(\mathbf{I}-\mathbf{A})^{-1}$ to both sides, first on one side and then on the other.

3.7.7. Use the result of Example 3.6.5 that says that $trace\,(\mathbf{AB})=trace\,(\mathbf{BA})$ to write

$$m=trace\,(\mathbf{I}_m)=trace\,(\mathbf{AB})=trace\,(\mathbf{BA})=trace\,(\mathbf{I}_n)=n.$$

3.7.8. Use the reverse order law for inversion to write

$$\left[\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\right]^{-1}=\mathbf{B}^{-1}(\mathbf{A}+\mathbf{B})\mathbf{A}^{-1}=\mathbf{B}^{-1}+\mathbf{A}^{-1}$$

and

$$\left[\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\right]^{-1}=\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})\mathbf{B}^{-1}=\mathbf{B}^{-1}+\mathbf{A}^{-1}.$$

3.7.9. (a) $(\mathbf{I}-\mathbf{S})\mathbf{x}=\mathbf{0} \implies \mathbf{x}^T(\mathbf{I}-\mathbf{S})\mathbf{x}=0 \implies \mathbf{x}^T\mathbf{x}=\mathbf{x}^T\mathbf{S}\mathbf{x}$. Taking transposes on both sides yields $\mathbf{x}^T\mathbf{x}=-\mathbf{x}^T\mathbf{S}\mathbf{x}$, so that $\mathbf{x}^T\mathbf{x}=0$, and thus $\mathbf{x}=\mathbf{0}$

(recall Exercise 3.6.12). The conclusion follows from property (3.7.8).
(b) First notice that Exercise 3.7.6 implies that $\mathbf{A} = (\mathbf{I}+\mathbf{S})(\mathbf{I}-\mathbf{S})^{-1} = (\mathbf{I}-\mathbf{S})^{-1}(\mathbf{I}+\mathbf{S})$. By using the reverse order laws, transposing both sides yields exactly the same thing as inverting both sides.

3.7.10. Use block multiplication to verify that the product of the matrix with its inverse is the identity matrix.

3.7.11. Use block multiplication to verify that the product of the matrix with its inverse is the identity matrix.

3.7.12. Let $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} \mathbf{D}^T & -\mathbf{B}^T \\ -\mathbf{C}^T & \mathbf{A}^T \end{pmatrix}$. The hypothesis implies that $\mathbf{MX} = \mathbf{I}$, and hence (from the discussion in Example 3.7.2) it must also be true that $\mathbf{XM} = \mathbf{I}$, from which the conclusion follows. **Note:** This problem appeared on a past Putnam Exam—a national mathematics competition for undergraduate students that is considered to be quite challenging. This means that you can be proud of yourself if you solved it before looking at this solution.

Solutions for exercises in section 3. 8

3.8.1. (a) $\mathbf{B}^{-1} = \begin{pmatrix} 1 & 2 & -1 \\ 0 & -1 & 1 \\ 1 & 4 & -2 \end{pmatrix}$

(b) Let $\mathbf{c} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ and $\mathbf{d}^T = \begin{pmatrix} 0 & 2 & 1 \end{pmatrix}$ to obtain $\mathbf{C}^{-1} = \begin{pmatrix} 0 & -2 & 1 \\ 1 & 3 & -1 \\ -1 & -4 & 2 \end{pmatrix}$

3.8.2. $\mathbf{A}_{*j}$ needs to be removed, and $\mathbf{b}$ needs to be inserted in its place. This is accomplished by writing $\mathbf{B} = \mathbf{A} + (\mathbf{b} - \mathbf{A}_{*j})\mathbf{e}_j^T$. Applying the Sherman–Morrison formula with $\mathbf{c} = \mathbf{b} - \mathbf{A}_{*j}$ and $\mathbf{d}^T = \mathbf{e}_j^T$ yields

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}(\mathbf{b}-\mathbf{A}_{*j})\mathbf{e}_j^T\mathbf{A}^{-1}}{1+\mathbf{e}_j^T\mathbf{A}^{-1}(\mathbf{b}-\mathbf{A}_{*j})} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{b}\mathbf{e}_j^T\mathbf{A}^{-1} - \mathbf{e}_j\mathbf{e}_j^T\mathbf{A}^{-1}}{1+\mathbf{e}_j^T\mathbf{A}^{-1}\mathbf{b} - \mathbf{e}_j^T\mathbf{e}_j}$$

$$= \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{b}[\mathbf{A}^{-1}]_{j*} - \mathbf{e}_j[\mathbf{A}^{-1}]_{j*}}{[\mathbf{A}^{-1}]_{j*}\mathbf{b}} = \mathbf{A}^{-1} - \frac{\big(\mathbf{A}^{-1}\mathbf{b} - \mathbf{e}_j\big)[\mathbf{A}^{-1}]_{j*}}{[\mathbf{A}^{-1}]_{j*}\mathbf{b}}.$$

3.8.3. Use the Sherman–Morrison formula to write

$$\mathbf{z} = (\mathbf{A}+\mathbf{c}\mathbf{d}^T)^{-1}\mathbf{b} = \left(\mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\mathbf{A}^{-1}}{1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}}\right)\mathbf{b} = \mathbf{A}^{-1}\mathbf{b} - \frac{\mathbf{A}^{-1}\mathbf{c}\mathbf{d}^T\mathbf{A}^{-1}\mathbf{b}}{1+\mathbf{d}^T\mathbf{A}^{-1}\mathbf{c}}$$

$$= \mathbf{x} - \frac{\mathbf{y}\mathbf{d}^T\mathbf{x}}{1+\mathbf{d}^T\mathbf{y}}.$$

3.8.4. (a) For a nonsingular matrix $\mathbf{A}$, the Sherman–Morrison formula guarantees that $\mathbf{A} + \alpha\mathbf{e}_i\mathbf{e}_j^T$ is also nonsingular when $1 + \alpha\left[\mathbf{A}^{-1}\right]_{ji} \neq 0$, and this certainly will be true if α is sufficiently small.

(b) Write $\mathbf{E}_{m\times m} = [\epsilon_{ij}] = \sum_{i,j=1}^{m} \epsilon_{ij}\mathbf{e}_i\mathbf{e}_j^T$ and successively apply part (a) to

$$\mathbf{I}+\mathbf{E} = \left(\left(\left(\mathbf{I}+\epsilon_{11}\mathbf{e}_1\mathbf{e}_1^T\right)+\epsilon_{12}\mathbf{e}_1\mathbf{e}_2^T\right)+\cdots+\epsilon_{mm}\mathbf{e}_m\mathbf{e}_m^T\right)$$

to conclude that when the ϵ_{ij} 's are sufficiently small,

$$\mathbf{I}+\epsilon_{11}\mathbf{e}_1\mathbf{e}_1^T, \quad \left(\left(\mathbf{I}+\epsilon_{11}\mathbf{e}_1\mathbf{e}_1^T\right)+\epsilon_{12}\mathbf{e}_1\mathbf{e}_2^T\right), \quad \ldots, \quad \mathbf{I}+\mathbf{E}$$

are each nonsingular.

3.8.5. Write $\mathbf{A}+\epsilon\mathbf{B} = \mathbf{A}(\mathbf{I}+\mathbf{A}^{-1}\epsilon\mathbf{B})$. You can either use the Neumann series result (3.8.5) or Exercise 3.8.4 to conclude that $(\mathbf{I}+\mathbf{A}^{-1}\epsilon\mathbf{B})$ is nonsingular whenever the entries of $\mathbf{A}^{-1}\epsilon\mathbf{B}$ are sufficiently small in magnitude, and this can be insured by restricting ϵ to a small enough interval about the origin. Since the product of two nonsingular matrices is again nonsingular—see (3.7.14)—it follows that $\mathbf{A}+\epsilon\mathbf{B} = \mathbf{A}(\mathbf{I}+\mathbf{A}^{-1}\epsilon\mathbf{B})$ must be nonsingular.

3.8.6. Since

$$\begin{pmatrix}\mathbf{I} & \mathbf{C}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{D}^T & -\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{D}^T & \mathbf{I}\end{pmatrix} = \begin{pmatrix}\mathbf{A}+\mathbf{C}\mathbf{D}^T & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}\end{pmatrix},$$

we can use $\mathbf{R} = \mathbf{D}^T$ and $\mathbf{B} = -\mathbf{I}$ in part (a) of Exercise 3.7.11 to obtain

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{D}^T & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1}+\mathbf{A}^{-1}\mathbf{C}\mathbf{S}^{-1}\mathbf{D}^T\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{C}\mathbf{S}^{-1}\\ -\mathbf{S}^{-1}\mathbf{D}^T\mathbf{A}^{-1} & \mathbf{S}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{I} & -\mathbf{C}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} = \begin{pmatrix}\left(\mathbf{A}+\mathbf{C}\mathbf{D}^T\right)^{-1} & \mathbf{0}\\ \mathbf{0} & -\mathbf{I}\end{pmatrix},$$

where $\mathbf{S} = -\left(\mathbf{I}+\mathbf{D}^T\mathbf{A}^{-1}\mathbf{C}\right)$. Comparing the upper-left-hand blocks produces

$$\left(\mathbf{A}+\mathbf{C}\mathbf{D}^T\right)^{-1} = \mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{C}\left(\mathbf{I}+\mathbf{D}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{D}^T\mathbf{A}^{-1}.$$

3.8.7. The ranking from best to worst condition is $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, because

$$\mathbf{A}^{-1} = \frac{1}{100}\begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 1\end{pmatrix} \implies \kappa(\mathbf{A}) = 20 = 2\times 10^1$$

$$\mathbf{B}^{-1} = \begin{pmatrix}-1465 & -161 & 17\\ 173 & 19 & -2\\ -82 & -9 & 1\end{pmatrix} \implies \kappa(\mathbf{B}) = 149,513 \approx 1.5\times 10^5$$

$$\mathbf{C}^{-1} = \begin{pmatrix}-42659 & 39794 & -948\\ 2025 & -1889 & 45\\ 45 & -42 & 1\end{pmatrix} \implies \kappa(\mathbf{C}) = 82,900,594 \approx 8.2\times 10^7.$$

3.8.8. (a) Differentiate $\mathbf{A}(t)\mathbf{A}(t)^{-1} = \mathbf{I}$ with the product rule for differentiation (recall Exercise 3.5.9).

(b) Use the product rule for differentiation together with part (a) to differentiate $\mathbf{A}(t)\mathbf{x}(t) = \mathbf{b}(t)$.

Solutions for exercises in section 3. 9

3.9.1. (a) If $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_k$ is the sequence of elementary matrices that corresponds to the elementary row operations used in the reduction $[\mathbf{A}|\mathbf{I}] \longrightarrow [\mathbf{B}|\mathbf{P}]$, then

$$\begin{aligned}\mathbf{G}_k \cdots \mathbf{G}_2\mathbf{G}_1[\mathbf{A}|\mathbf{I}] = [\mathbf{B}|\mathbf{P}] \quad &\Longrightarrow \quad [\mathbf{G}_k \cdots \mathbf{G}_2\mathbf{G}_1\mathbf{A} \mid \mathbf{G}_k \cdots \mathbf{G}_2\mathbf{G}_1\mathbf{I}] = [\mathbf{B}|\mathbf{P}] \\ &\Longrightarrow \quad \mathbf{G}_k \cdots \mathbf{G}_2\mathbf{G}_1\mathbf{A} = \mathbf{B} \quad \text{and} \quad \mathbf{G}_k \cdots \mathbf{G}_2\mathbf{G}_1 = \mathbf{P}.\end{aligned}$$

(b) Use the same argument given above, but apply it on the right-hand side.

(c) $[\mathbf{A}|\mathbf{I}] \xrightarrow{\text{Gauss–Jordan}} [\mathbf{E_A}|\mathbf{P}]$ yields

$$\left(\begin{array}{cccc|ccc} 1 & 2 & 3 & 4 & 1 & 0 & 0 \\ 2 & 4 & 6 & 7 & 0 & 1 & 0 \\ 1 & 2 & 3 & 6 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{cccc|ccc} 1 & 2 & 3 & 0 & -7 & 4 & 0 \\ 0 & 0 & 0 & 1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -5 & 2 & 1 \end{array}\right).$$

Thus $\mathbf{P} = \begin{pmatrix} -7 & 4 & 0 \\ 2 & -1 & 0 \\ -5 & 2 & 1 \end{pmatrix}$ is the product of the elementary matrices corresponding to the operations used in the reduction, and $\mathbf{PA} = \mathbf{E_A}$.

(d) You already have $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{E_A}$. Now find $\mathbf{Q}$ such that $\mathbf{E_A Q} = \mathbf{N}_r$ by column reducing $\mathbf{E_A}$. Proceed using part (b) to accumulate $\mathbf{Q}$.

$$\left[\begin{array}{c} \mathbf{E_A} \\ \hline \mathbf{I}_4 \end{array}\right] \longrightarrow \left(\begin{array}{cccc} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{cccc} 1 & 0 & 2 & 3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right) \longrightarrow \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 0 & -2 & -3 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right)$$

3.9.2. (a) Yes—because $rank\,(\mathbf{A}) = rank\,(\mathbf{B})$. (b) Yes—because $\mathbf{E_A} = \mathbf{E_B}$. (c) No—because $\mathbf{E}_{\mathbf{A}^T} \neq \mathbf{E}_{\mathbf{B}^T}$.

3.9.3. The positions of the basic columns in $\mathbf{A}$ correspond to those in $\mathbf{E_A}$. Because $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \Longleftrightarrow \mathbf{E_A} = \mathbf{E_B}$, it follows that the basic columns in $\mathbf{A}$ and $\mathbf{B}$ must be in the same positions.

3.9.4. An elementary interchange matrix (a Type I matrix) has the form $\mathbf{E} = \mathbf{I} - \mathbf{u}\mathbf{u}^T$, where $\mathbf{u} = \mathbf{e}_i - \mathbf{e}_j$, and it follows from (3.9.1) that $\mathbf{E} = \mathbf{E}^T = \mathbf{E}^{-1}$. If $\mathbf{P} = \mathbf{E}_1\mathbf{E}_2 \cdots \mathbf{E}_k$ is a product of elementary interchange matrices, then the reverse order laws yield

$$\begin{aligned}\mathbf{P}^{-1} &= (\mathbf{E}_1\mathbf{E}_2 \cdots \mathbf{E}_k)^{-1} = \mathbf{E}_k^{-1} \cdots \mathbf{E}_2^{-1}\mathbf{E}_1^{-1} \\ &= \mathbf{E}_k^T \cdots \mathbf{E}_2^T\mathbf{E}_1^T = (\mathbf{E}_1\mathbf{E}_2 \cdots \mathbf{E}_k)^T = \mathbf{P}^T.\end{aligned}$$

3.9.5. They are all true! $\mathbf{A} \sim \mathbf{I} \sim \mathbf{A}^{-1}$ because $rank\,(\mathbf{A}) = n = rank\left(\mathbf{A}^{-1}\right)$, $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{A}^{-1}$ because $\mathbf{PA} = \mathbf{A}^{-1}$ with $\mathbf{P} = \left(\mathbf{A}^{-1}\right)^2 = \mathbf{A}^{-2}$, and $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{A}^{-1}$ because $\mathbf{AQ} = \mathbf{A}^{-1}$ with $\mathbf{Q} = \mathbf{A}^{-2}$. The fact that $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{I}$ and $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{I}$ follows since $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

3.9.6. (a), (c), (d), and (e) are true.

3.9.7. Rows i and j can be interchanged with the following sequence of Type II and Type III operations—this is Exercise 1.2.12 on p. 14.

$$\begin{aligned}
R_j &\leftarrow R_j + R_i && \text{(replace row } j \text{ by the sum of row } j \text{ and } i\text{)}\\
R_i &\leftarrow R_i - R_j && \text{(replace row } i \text{ by the difference of row } i \text{ and } j\text{)}\\
R_j &\leftarrow R_j + R_i && \text{(replace row } j \text{ by the sum of row } j \text{ and } i\text{)}\\
R_i &\leftarrow -R_i && \text{(replace row } i \text{ by its negative)}
\end{aligned}$$

Translating these to elementary matrices (remembering to build from the right to the left) produces

$$(\mathbf{I} - 2\mathbf{e}_i\mathbf{e}_i^T)(\mathbf{I} + \mathbf{e}_j\mathbf{e}_i^T)(\mathbf{I} - \mathbf{e}_i\mathbf{e}_j^T)(\mathbf{I} + \mathbf{e}_j\mathbf{e}_i^T) = \mathbf{I} - \mathbf{u}\mathbf{u}^T, \quad \text{where} \quad \mathbf{u} = \mathbf{e}_i - \mathbf{e}_j.$$

3.9.8. Let $\mathbf{B}_{m\times r} = [\mathbf{A}_{*b_1}\mathbf{A}_{*b_2}\cdots\mathbf{A}_{*b_r}]$ contain the basic columns of $\mathbf{A}$, and let $\mathbf{C}_{r\times n}$ contain the nonzero rows of $\mathbf{E}_{\mathbf{A}}$. If $\mathbf{A}_{*k}$ is basic—say $\mathbf{A}_{*k} = \mathbf{A}_{*b_j}$—then $\mathbf{C}_{*k} = \mathbf{e}_j$, and

$$(\mathbf{BC})_{*k} = \mathbf{BC}_{*k} = \mathbf{Be}_j = \mathbf{B}_{*j} = \mathbf{A}_{*b_j} = \mathbf{A}_{*k}.$$

If $\mathbf{A}_{*k}$ is nonbasic, then $\mathbf{C}_{*k}$ is nonbasic and has the form

$$\begin{aligned}
\mathbf{C}_{*k} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_j \\ \vdots \\ 0 \end{pmatrix} &= \mu_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \mu_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \mu_j \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}\\
&= \mu_1\mathbf{e}_1 + \mu_2\mathbf{e}_2 + \cdots + \mu_j\mathbf{e}_j,
\end{aligned}$$

where the $\mathbf{e}_i$'s are the basic columns to the left of $\mathbf{C}_{*k}$. Because $\mathbf{A} \overset{\text{row}}{\sim} \mathbf{E}_{\mathbf{A}}$, the relationships that exist among the columns of $\mathbf{A}$ are exactly the same as the relationships that exist among the columns of $\mathbf{E}_{\mathbf{A}}$. In particular,

$$\mathbf{A}_{*k} = \mu_1\mathbf{A}_{*b_1} + \mu_2\mathbf{A}_{*b_2} + \cdots + \mu_j\mathbf{A}_{*b_j},$$

where the $\mathbf{A}_{*b_i}$'s are the basic columns to the left of $\mathbf{A}_{*k}$. Therefore,

$$\begin{aligned}
(\mathbf{BC})_{*k} = \mathbf{BC}_{*k} &= \mathbf{B}\left(\mu_1\mathbf{e}_1 + \mu_2\mathbf{e}_2 + \cdots + \mu_j\mathbf{e}_j\right)\\
&= \mu_1\mathbf{B}_{*1} + \mu_2\mathbf{B}_{*2} + \cdots + \mu_j\mathbf{B}_{*j}\\
&= \mu_1\mathbf{A}_{*b_1} + \mu_2\mathbf{A}_{*b_2} + \cdots + \mu_j\mathbf{A}_{*b_j}\\
&= \mathbf{A}_{*k}.
\end{aligned}$$

3.9.9. If $\mathbf{A} = \mathbf{u}\mathbf{v}^T$, where $\mathbf{u}_{m\times 1}$ and $\mathbf{v}_{n\times 1}$ are nonzero columns, then

$$\mathbf{u} \overset{\text{row}}{\sim} \mathbf{e}_1 \quad \text{and} \quad \mathbf{v}^T \overset{\text{col}}{\sim} \mathbf{e}_1^T \implies \mathbf{A} = \mathbf{u}\mathbf{v}^T \sim \mathbf{e}_1\mathbf{e}_1^T = \mathbf{N}_1 \implies rank\,(\mathbf{A}) = 1.$$

Conversely, if $rank\,(\mathbf{A}) = 1$, then the existence of $\mathbf{u}$ and $\mathbf{v}$ follows from Exercise 3.9.8. If you do not wish to rely on Exercise 3.9.8, write $\mathbf{PAQ} = \mathbf{N}_1 = \mathbf{e}_1\mathbf{e}_1^T$, where $\mathbf{e}_1$ is $m \times 1$ and $\mathbf{e}_1^T$ is $1 \times n$ so that

$$\mathbf{A} = \mathbf{P}^{-1}\mathbf{e}_1\mathbf{e}_1^T\mathbf{Q}^{-1} = \left(\mathbf{P}^{-1}\right)_{*1}\left(\mathbf{Q}^{-1}\right)_{1*} = \mathbf{u}\mathbf{v}^T.$$

3.9.10. Use Exercise 3.9.9 and write

$$\mathbf{A} = \mathbf{u}\mathbf{v}^T \implies \mathbf{A}^2 = \left(\mathbf{u}\mathbf{v}^T\right)\left(\mathbf{u}\mathbf{v}^T\right) = \mathbf{u}\left(\mathbf{v}^T\mathbf{u}\right)\mathbf{v}^T = \tau\mathbf{u}\mathbf{v}^T = \tau\mathbf{A},$$

where $\tau = \mathbf{v}^T\mathbf{u}$. Recall from Example 3.6.5 that $trace\,(\mathbf{AB}) = trace\,(\mathbf{BA})$, and write

$$\tau = trace(\tau) = trace\left(\mathbf{v}^T\mathbf{u}\right) = trace\left(\mathbf{u}\mathbf{v}^T\right) = trace\,(\mathbf{A}).$$

Solutions for exercises in section 3. 10

3.10.1. (a) $\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} 1 & 4 & 5 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{pmatrix}$ (b) $\mathbf{x}_1 = \begin{pmatrix} 110 \\ -36 \\ 8 \end{pmatrix}$ and $\mathbf{x}_2 = \begin{pmatrix} 112 \\ -39 \\ 10 \end{pmatrix}$

(c) $\mathbf{A}^{-1} = \frac{1}{6}\begin{pmatrix} 124 & -40 & 14 \\ -42 & 15 & -6 \\ 10 & -4 & 2 \end{pmatrix}$

3.10.2. (a) The second pivot is zero. (b) $\mathbf{P}$ is the permutation matrix associated with the permutation $\mathbf{p} = (\,2 \quad 4 \quad 1 \quad 3\,)$. $\mathbf{P}$ is constructed by permuting the rows of $\mathbf{I}$ in this manner.

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \\ 2/3 & -1/2 & 1/2 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 3 & 6 & -12 & 3 \\ 0 & 2 & -2 & 6 \\ 0 & 0 & 8 & 16 \\ 0 & 0 & 0 & -5 \end{pmatrix}$$

(c) $\mathbf{x} = \begin{pmatrix} 2 \\ -1 \\ 0 \\ 1 \end{pmatrix}$

3.10.3. $\xi = 0,\ \pm\sqrt{2},\ \pm\sqrt{3}$

3.10.4. $\mathbf{A}$ possesses an LU factorization if and only if all leading principal submatrices are nonsingular. The argument associated with equation (3.10.13) proves that

$$\begin{pmatrix} \mathbf{L}_k & \mathbf{0} \\ \mathbf{c}^T\mathbf{U}_k^{-1} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{U}_k & \mathbf{L}_k^{-1}\mathbf{b} \\ \mathbf{0} & a_{k+1,k+1} - \mathbf{c}^T\mathbf{A}_k^{-1}\mathbf{b} \end{pmatrix} = \mathbf{L}_{k+1}\mathbf{U}_{k+1}$$

is the LU factorization for $\mathbf{A}_{k+1}$. The desired conclusion follows from the fact that the $k+1^{th}$ pivot is the $(k+1, k+1)$-entry in $\mathbf{U}_{k+1}$. This pivot must be nonzero because $\mathbf{U}_{k+1}$ is nonsingular.

3.10.5. If $\mathbf{L}$ and $\mathbf{U}$ are both triangular with 1's on the diagonal, then $\mathbf{L}^{-1}$ and $\mathbf{U}^{-1}$ contain only integer entries, and consequently $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$ is an integer matrix.

3.10.6. (b) $\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ 0 & -2/3 & 1 & 0 \\ 0 & 0 & -3/4 & 1 \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & 3/2 & -1 & 0 \\ 0 & 0 & 4/3 & -1 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}$

3.10.7. Observe how the LU factors evolve from Gaussian elimination. Following the procedure described in Example 3.10.1 where multipliers ℓ_{ij} are stored in the positions they annihilate (i.e., in the (i,j)-position), and where $\star$'s are put in positions that can be nonzero, the reduction of a 5×5 band matrix with bandwidth $w = 2$ proceeds as shown below.

$$\begin{pmatrix} \star & \star & \star & 0 & 0 \\ \star & \star & \star & \star & 0 \\ \star & \star & \star & \star & \star \\ 0 & \star & \star & \star & \star \\ 0 & 0 & \star & \star & \star \end{pmatrix} \longrightarrow \begin{pmatrix} \star & \star & \star & 0 & 0 \\ l_{21} & \star & \star & \star & 0 \\ l_{31} & \star & \star & \star & \star \\ 0 & \star & \star & \star & \star \\ 0 & 0 & \star & \star & \star \end{pmatrix} \longrightarrow \begin{pmatrix} \star & \star & \star & 0 & 0 \\ l_{21} & \star & \star & \star & 0 \\ l_{31} & l_{32} & \star & \star & \star \\ 0 & l_{42} & \star & \star & \star \\ 0 & 0 & \star & \star & \star \end{pmatrix}$$

$$\longrightarrow \begin{pmatrix} \star & \star & \star & 0 & 0 \\ l_{21} & \star & \star & \star & 0 \\ l_{31} & l_{32} & \star & \star & \star \\ 0 & l_{42} & l_{43} & \star & \star \\ 0 & 0 & l_{53} & \star & \star \end{pmatrix} \longrightarrow \begin{pmatrix} \star & \star & \star & 0 & 0 \\ l_{21} & \star & \star & \star & 0 \\ l_{31} & l_{32} & \star & \star & \star \\ 0 & l_{42} & l_{43} & \star & \star \\ 0 & 0 & l_{53} & l_{54} & \star \end{pmatrix}$$

Thus $\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 & 0 \\ 0 & l_{42} & l_{43} & 1 & 0 \\ 0 & 0 & l_{53} & l_{54} & 1 \end{pmatrix}$ and $\mathbf{U} = \begin{pmatrix} \star & \star & \star & 0 & 0 \\ 0 & \star & \star & \star & 0 \\ 0 & 0 & \star & \star & \star \\ 0 & 0 & 0 & \star & \star \\ 0 & 0 & 0 & 0 & \star \end{pmatrix}$.

3.10.8. (a) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (b) $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

3.10.9. (a) $\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 4 & 1 & 0 \\ 3 & 2 & 1 \end{pmatrix}$, $\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, and $\mathbf{U} = \begin{pmatrix} 1 & 4 & 5 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$

(b) Use the same argument given for the uniqueness of the LU factorization with minor modifications.

(c) $\mathbf{A} = \mathbf{A}^T \implies \mathbf{LDU} = \mathbf{U}^T\mathbf{D}^T\mathbf{L}^T = \mathbf{U}^T\mathbf{D}\mathbf{L}^T$. These are each LDU factorizations for $\mathbf{A}$, and consequently the uniqueness of the LDU factorization means that $\mathbf{U} = \mathbf{L}^T$.

3.10.10. $\mathbf{A}$ is symmetric with pivots 1, 4, 9. The Cholesky factor is $\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 3 & 3 \end{pmatrix}$.

It is unworthy of excellent men to lose hours
like slaves in the labor of calculations.
— *Baron Gottfried Wilhelm von Leibnitz* (1646–1716)

Solutions for Chapter 4

Solutions for exercises in section 4. 1

4.1.1. Only (b) and (d) are subspaces.

4.1.2. (a), (b), (f), (g), and (i) are subspaces.

4.1.3. All of $\Re^3$.

4.1.4. If $\mathbf{v} \in \mathcal{V}$ is a nonzero vector in a space $\mathcal{V}$, then all scalar multiples $\alpha\mathbf{v}$ must also be in $\mathcal{V}$.

4.1.5. (a) A line. (b) The (x,y)-plane. (c) $\Re^3$

4.1.6. Only (c) and (e) span $\Re^3$. To see that (d) does not span $\Re^3$, ask whether or not every vector $(x, y, z) \in \Re^3$ can be written as a linear combination of the vectors in (d). It's convenient to think in terms columns, so rephrase the question by asking if every $\mathbf{b} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ can be written as a linear combination of $\left\{ \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 4 \\ 4 \\ 1 \end{pmatrix} \right\}$. That is, for each $\mathbf{b} \in \Re^3$, are there scalars $\alpha_1, \alpha_2, \alpha_3$ such that $\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \alpha_3\mathbf{v}_3 = \mathbf{b}$ or, equivalently, is

$$\begin{pmatrix} 1 & 2 & 4 \\ 2 & 0 & 4 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{consistent for all} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix}?$$

This is a system of the form $\mathbf{Ax} = \mathbf{b}$, and it is consistent for all $\mathbf{b}$ if and only if $rank\,([\mathbf{A}|\mathbf{b}]) = rank\,(\mathbf{A})$ for all $\mathbf{b}$. Since

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

it's clear that there exist $\mathbf{b}$'s (e.g., $\mathbf{b} = (1, 0, 0)^T$) for which $\mathbf{Ax} = \mathbf{b}$ is not consistent, and hence not all $\mathbf{b}$'s are a combination of the $\mathbf{v}_i$'s. Therefore, the $\mathbf{v}_i$'s don't span $\Re^3$.

4.1.7. This follows from (4.1.2).

4.1.8. (a) $\mathbf{u}, \mathbf{v} \in \mathcal{X} \cap \mathcal{Y} \implies \mathbf{u}, \mathbf{v} \in \mathcal{X}$ and $\mathbf{u}, \mathbf{v} \in \mathcal{Y}$. Because $\mathcal{X}$ and $\mathcal{Y}$ are closed with respect to addition, it follows that $\mathbf{u}+\mathbf{v} \in \mathcal{X}$ and $\mathbf{u}+\mathbf{v} \in \mathcal{Y}$, and therefore $\mathbf{u}+\mathbf{v} \in \mathcal{X} \cap \mathcal{Y}$. Because $\mathcal{X}$ and $\mathcal{Y}$ are both closed with respect to scalar multiplication, we have that $\alpha\mathbf{u} \in \mathcal{X}$ and $\alpha\mathbf{u} \in \mathcal{Y}$ for all α, and consequently $\alpha\mathbf{u} \in \mathcal{X} \cap \mathcal{Y}$ for all α.

(b) The union of two different lines through the origin in $\Re^2$ is not a subspace.

4.1.9. (a) **(A1)** holds because $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{A}(\mathcal{S}) \implies \mathbf{x}_1 = \mathbf{A}\mathbf{s}_1$ and $\mathbf{x}_2 = \mathbf{A}\mathbf{s}_2$ for some $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S} \implies \mathbf{x}_1 + \mathbf{x}_2 = \mathbf{A}(\mathbf{s}_1 + \mathbf{s}_2)$. Since $\mathcal{S}$ is a subspace, it is closed under vector addition, so $\mathbf{s}_1 + \mathbf{s}_2 \in \mathcal{S}$. Therefore, $\mathbf{x}_1 + \mathbf{x}_2$ is the image of something in $\mathcal{S}$—namely, $\mathbf{s}_1 + \mathbf{s}_2$—and this means that $\mathbf{x}_1 + \mathbf{x}_2 \in \mathbf{A}(\mathcal{S})$. To see that **(M1)** holds, consider $\alpha\mathbf{x}$, where α is an arbitrary scalar and $\mathbf{x} \in \mathbf{A}(\mathcal{S})$. Now, $\mathbf{x} \in \mathbf{A}(\mathcal{S}) \implies \mathbf{x} = \mathbf{A}\mathbf{s}$ for some $\mathbf{s} \in \mathcal{S} \implies \alpha\mathbf{x} = \alpha\mathbf{A}\mathbf{s} = \mathbf{A}(\alpha\mathbf{s})$. Since $\mathcal{S}$ is a subspace, we are guaranteed that $\alpha\mathbf{s} \in \mathcal{S}$, and therefore $\alpha\mathbf{x}$ is the image of something in $\mathcal{S}$. This is what it means to say $\alpha\mathbf{x} \in \mathbf{A}(\mathcal{S})$.

(b) Prove equality by demonstrating that $span\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\} \subseteq \mathbf{A}(\mathcal{S})$ and $\mathbf{A}(\mathcal{S}) \subseteq span\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\}$. To show $span\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\} \subseteq \mathbf{A}(\mathcal{S})$, write

$$\mathbf{x} \in span\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\} \implies \mathbf{x} = \sum_{i=1}^{k} \alpha_i(\mathbf{A}\mathbf{s}_i) = \mathbf{A}\left(\sum_{i=1}^{k} \alpha_i\mathbf{s}_i\right) \in \mathbf{A}(\mathcal{S}).$$

Inclusion in the reverse direction is established by saying

$$\mathbf{x} \in \mathbf{A}(\mathcal{S}) \implies \mathbf{x} = \mathbf{A}\mathbf{s} \text{ for some } \mathbf{s} \in \mathcal{S} \implies \mathbf{s} = \sum_{i=1}^{k} \beta_i\mathbf{s}_i$$

$$\implies \mathbf{x} = \mathbf{A}\left(\sum_{i=1}^{k} \beta_i\mathbf{s}_i\right) = \sum_{i=1}^{k} \beta_i(\mathbf{A}\mathbf{s}_i) \in span\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\}.$$

4.1.10. (a) Yes, all of the defining properties are satisfied.

(b) Yes, this is essentially $\Re^2$.

(c) No, it is not closed with respect to scalar multiplication.

4.1.11. If $span(\mathcal{M}) = span(\mathcal{N})$, then every vector in $\mathcal{N}$ must be a linear combination of vectors from $\mathcal{M}$. In particular, $\mathbf{v}$ must be a linear combination of the $\mathbf{m}_i$'s, and hence $\mathbf{v} \in span(\mathcal{M})$. To prove the converse, first notice that $span(\mathcal{M}) \subseteq span(\mathcal{N})$. The desired conclusion will follow if it can be demonstrated that $span(\mathcal{M}) \supseteq span(\mathcal{N})$. The hypothesis that $\mathbf{v} \in span(\mathcal{M})$ guarantees that $\mathbf{v} = \sum_{i=1}^{r} \beta_i\mathbf{m}_i$. If $\mathbf{z} \in span(\mathcal{N})$, then

$$\begin{aligned}\mathbf{z} &= \sum_{i=1}^{r} \alpha_i\mathbf{m}_i + \alpha_{r+1}\mathbf{v} = \sum_{i=1}^{r} \alpha_i\mathbf{m}_i + \alpha_{r+1}\sum_{i=1}^{r} \beta_i\mathbf{m}_i \\ &= \sum_{i=1}^{r} \Big(\alpha_i + \alpha_{r+1}\beta_i\Big)\mathbf{m}_i,\end{aligned}$$

which shows $\mathbf{z} \in span\,(\mathcal{M})$, and therefore $span\,(\mathcal{M}) \supseteq span\,(\mathcal{N})$.

4.1.12. To show $span\,(\mathcal{S}) \subseteq \mathcal{M}$, observe that $\mathbf{x} \in span\,(\mathcal{S}) \implies \mathbf{x} = \sum_i \alpha_i \mathbf{v}_i$. If $\mathcal{V}$ is any subspace containing $\mathcal{S}$, then $\sum_i \alpha_i \mathbf{v}_i \in \mathcal{V}$ because $\mathcal{V}$ is closed under addition and scalar multiplication, and therefore $\mathbf{x} \in \mathcal{M}$. The fact that $\mathcal{M} \subseteq span\,(\mathcal{S})$ follows because if $\mathbf{x} \in \mathcal{M}$, then $\mathbf{x} \in span\,(\mathcal{S})$ because $span\,(\mathcal{S})$ is one particular subspace that contains $\mathcal{S}$.

Solutions for exercises in section 4. 2

4.2.1. $R\,(\mathbf{A}) = span\left\{\begin{pmatrix}1\\-2\\1\end{pmatrix}, \begin{pmatrix}1\\0\\2\end{pmatrix}\right\}, \quad N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix}4\\1\\-2\end{pmatrix}\right\},$

$$N\,(\mathbf{A}) = span\left\{\begin{pmatrix}-2\\1\\0\\0\\0\end{pmatrix}, \begin{pmatrix}2\\0\\-3\\1\\0\end{pmatrix}, \begin{pmatrix}-1\\0\\-4\\0\\1\end{pmatrix}\right\},$$

$$R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix}1\\2\\0\\-2\\1\end{pmatrix}, \begin{pmatrix}0\\0\\1\\3\\4\end{pmatrix}\right\}.$$

4.2.2. (a) This is simply a restatement of equation (4.2.3).
(b) $\mathbf{Ax} = \mathbf{b}$ has a unique solution if and only if $rank\,(\mathbf{A}) = n$ (i.e., there are no free variables—see §2.5), and (4.2.10) says $rank\,(\mathbf{A}) = n \iff N\,(\mathbf{A}) = \{\mathbf{0}\}$.

4.2.3. (a) It is consistent because $\mathbf{b} \in R\,(\mathbf{A})$.
(b) It is nonunique because $N\,(\mathbf{A}) \neq \{\mathbf{0}\}$—see Exercise 4.2.2.

4.2.4. Yes, because $rank[\mathbf{A}|\mathbf{b}] = rank\,(\mathbf{A}) = 3 \implies \exists\ \mathbf{x}$ such that $\mathbf{Ax} = \mathbf{b}$—i.e., $\mathbf{Ax} = \mathbf{b}$ is consistent.

4.2.5. (a) If $R\,(\mathbf{A}) = \Re^n$, then

$$R\,(\mathbf{A}) = R\,(\mathbf{I}_n) \implies \mathbf{A} \overset{\text{col}}{\sim} \mathbf{I}_n \implies rank\,(\mathbf{A}) = rank\,(\mathbf{I}_n) = n.$$

(b) $R\,(\mathbf{A}) = R\left(\mathbf{A}^T\right) = \Re^n$ and $N\,(\mathbf{A}) = N\left(\mathbf{A}^T\right) = \{\mathbf{0}\}$.

4.2.6. $\mathbf{E_A} \neq \mathbf{E_B}$ means that $R\left(\mathbf{A}^T\right) \neq R\left(\mathbf{B}^T\right)$ and $N\,(\mathbf{A}) \neq N\,(\mathbf{B})$. However, $\mathbf{E}_{\mathbf{A}^T} = \mathbf{E}_{\mathbf{B}^T}$ implies that $R\,(\mathbf{A}) = R\,(\mathbf{B})$ and $N\left(\mathbf{A}^T\right) = N\left(\mathbf{B}^T\right)$.

4.2.7. Demonstrate that $rank\,(\mathbf{A}_{n\times n}) = n$ by using (4.2.10). If $\mathbf{x} \in N\,(\mathbf{A})$, then

$$\begin{aligned}
\mathbf{Ax} = \mathbf{0} &\implies \mathbf{A}_1\mathbf{x} = \mathbf{0} \quad \text{and} \quad \mathbf{A}_2\mathbf{x} = \mathbf{0} \\
&\implies \mathbf{x} \in N\,(\mathbf{A}_1) = R\left(\mathbf{A}_2^T\right) \implies \exists\ \mathbf{y}^T \text{ such that } \mathbf{x}^T = \mathbf{y}^T\mathbf{A}_2 \\
&\implies \mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}_2\mathbf{x} = \mathbf{0} \implies \sum_i x_i^2 = 0 \implies \mathbf{x} = \mathbf{0}.
\end{aligned}$$

4.2.8. $\mathbf{y}^T\mathbf{b} = 0 \;\forall\; \mathbf{y} \in N\left(\mathbf{A}^T\right) = R\left(\mathbf{P}_2^T\right) \implies \mathbf{P}_2\mathbf{b} = 0 \implies \mathbf{b} \in N\left(\mathbf{P}_2\right) = R\left(\mathbf{A}\right)$

4.2.9. $\mathbf{x} \in R\big(\mathbf{A} \mid \mathbf{B}\big) \iff \exists\; \mathbf{y}$ such that $\mathbf{x} = \big(\mathbf{A} \mid \mathbf{B}\big)\mathbf{y} = \big(\mathbf{A} \mid \mathbf{B}\big)\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} = \mathbf{A}\mathbf{y}_1 + \mathbf{B}\mathbf{y}_2 \iff \mathbf{x} \in R\left(\mathbf{A}\right) + R\left(\mathbf{B}\right)$

4.2.10. (a) $\mathbf{p} + N\left(\mathbf{A}\right)$ is the set of all possible solutions to $\mathbf{A}\mathbf{x} = \mathbf{b}$. Recall from (2.5.7) that the general solution of a nonhomogeneous equation is a particular solution plus the general solution of the homogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{0}$. The general solution of the homogeneous equation is simply a way of describing all possible solutions of $\mathbf{A}\mathbf{x} = \mathbf{0}$, which is $N\left(\mathbf{A}\right)$.
(b) $rank\left(\mathbf{A}_{3\times 3}\right) = 1$ means that $N\left(\mathbf{A}\right)$ is spanned by two vectors, and hence $N\left(\mathbf{A}\right)$ is a plane through the origin. From the parallelogram law, $\mathbf{p} + N\left(\mathbf{A}\right)$ is a plane parallel to $N\left(\mathbf{A}\right)$ passing through the point defined by $\mathbf{p}$.
(c) This time $N\left(\mathbf{A}\right)$ is spanned by a single vector, and $\mathbf{p} + N\left(\mathbf{A}\right)$ is a line parallel to $N\left(\mathbf{A}\right)$ passing through the point defined by $\mathbf{p}$.

4.2.11. $\mathbf{a} \in R\left(\mathbf{A}^T\right) \iff \exists\; \mathbf{y}$ such that $\mathbf{a}^T = \mathbf{y}^T\mathbf{A}$. If $\mathbf{A}\mathbf{x} = \mathbf{b}$, then

$$\mathbf{a}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{b},$$

which is independent of $\mathbf{x}$.

4.2.12. (a) $\mathbf{b} \in R\left(\mathbf{A}\mathbf{B}\right) \implies \exists\; \mathbf{x}$ such that $\mathbf{b} = \mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{x}) \implies \mathbf{b} \in R\left(\mathbf{A}\right)$ because $\mathbf{b}$ is the image of $\mathbf{B}\mathbf{x}$.
(b) $\mathbf{x} \in N\left(\mathbf{B}\right) \implies \mathbf{B}\mathbf{x} = \mathbf{0} \implies \mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{0} \implies \mathbf{x} \in N\left(\mathbf{A}\mathbf{B}\right)$.

4.2.13. Given any $\mathbf{z} \in R\left(\mathbf{A}\mathbf{B}\right)$, the object is to show that $\mathbf{z}$ can be written as some linear combination of the $\mathbf{A}\mathbf{b}_i$'s. Argue as follows. $\mathbf{z} \in R\left(\mathbf{A}\mathbf{B}\right) \implies \mathbf{z} = \mathbf{A}\mathbf{B}\mathbf{y}$ for some $\mathbf{y}$. But it is always true that $\mathbf{B}\mathbf{y} \in R\left(\mathbf{B}\right)$, so

$$\mathbf{B}\mathbf{y} = \alpha_1\mathbf{b}_1 + \alpha_2\mathbf{b}_2 + \cdots + \alpha_n\mathbf{b}_n,$$

and therefore $\mathbf{z} = \mathbf{A}\mathbf{B}\mathbf{y} = \alpha_1\mathbf{A}\mathbf{b}_1 + \alpha_2\mathbf{A}\mathbf{b}_2 + \cdots + \alpha_n\mathbf{A}\mathbf{b}_n$.

Solutions for exercises in section 4. 3

4.3.1. (a) and (b) are linearly dependent—all others are linearly independent. To write one vector as a combination of others in a dependent set, place the vectors as columns in $\mathbf{A}$ and find $\mathbf{E}_{\mathbf{A}}$. This reveals the dependence relationships among columns of $\mathbf{A}$.

4.3.2. (a) According to (4.3.12), the basic columns in $\mathbf{A}$ always constitute one maximal linearly independent subset.
(b) Ten—5 sets using two vectors, 4 sets using one vector, and the empty set.

4.3.3. $rank\left(\mathbf{H}\right) \leq 3$, and according to (4.3.11), $rank\left(\mathbf{H}\right)$ is the maximal number of independent rows in $\mathbf{H}$.

4.3.4. The question is really whether or not the columns in

$$\hat{\mathbf{A}} = \begin{array}{c} \\ \#1 \\ \#2 \\ \#3 \\ \#4 \end{array} \begin{array}{c} \begin{array}{cccc} & S & L & F \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 10 \\ 1 & 2 & 1 & 12 \\ 1 & 2 & 2 & 15 \\ 1 & 3 & 2 & 17 \end{pmatrix} \end{array}$$

are linearly independent. Reducing $\hat{\mathbf{A}}$ to $\mathbf{E}_{\hat{\mathbf{A}}}$ shows that $5+2S+3L-F=0$.

4.3.5. (a) This follows directly from the definition of linear dependence because there are nonzero values of α such that $\alpha\mathbf{0}=\mathbf{0}$.

(b) This is a consequence of (4.3.13).

4.3.6. If each $t_{ii} \neq 0$, then $\mathbf{T}$ is nonsingular, and the result follows from (4.3.6) and (4.3.7).

4.3.7. It is linearly independent because

$$\alpha_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} + \alpha_4 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\iff \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \alpha_4 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\iff \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \iff \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

4.3.8. $\mathbf{A}$ is nonsingular because it is diagonally dominant.

4.3.9. $\mathcal{S}$ is linearly independent using exact arithmetic, but using 3-digit arithmetic yields the conclusion that $\mathcal{S}$ is dependent.

4.3.10. If $\mathbf{e}$ is the column vector of all 1's, then $\mathbf{Ae}=\mathbf{0}$, so that $N(\mathbf{A}) \neq \{\mathbf{0}\}$.

4.3.11. (Solution 1.) $\sum_i \alpha_i \mathbf{P}\mathbf{u}_i = \mathbf{0} \implies \mathbf{P}\sum_i \alpha_i \mathbf{u}_i = \mathbf{0} \implies \sum_i \alpha_i \mathbf{u}_i = \mathbf{0} \implies$ each $\alpha_i = 0$ because the $\mathbf{u}_i$'s are linearly independent.

(Solution 2.) If $\mathbf{A}_{m\times n}$ is the matrix containing the $\mathbf{u}_i$'s as columns, then $\mathbf{PA}=\mathbf{B}$ is the matrix containing the vectors in $\mathbf{P}(\mathcal{S})$ as its columns. Now,

$$\mathbf{A} \overset{\text{row}}{\sim} \mathbf{B} \implies rank\,(\mathbf{B}) = rank\,(\mathbf{A}) = n,$$

and hence (4.3.3) insures that the columns of $\mathbf{B}$ are linearly independent. The result need not be true if $\mathbf{P}$ is singular—take $\mathbf{P}=\mathbf{0}$ for example.

4.3.12. If $\mathbf{A}_{m\times n}$ is the matrix containing the $\mathbf{u}_i$'s as columns, and if

$$\mathbf{Q}_{n\times n} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

then the columns of $\mathbf{B} = \mathbf{AQ}$ are the vectors in $\mathcal{S}'$. Clearly, $\mathbf{Q}$ is nonsingular so that $\mathbf{A} \overset{\text{col}}{\sim} \mathbf{B}$, and thus $rank\,(\mathbf{A}) = rank\,(\mathbf{B})$. The desired result now follows from (4.3.3).

4.3.13. (a) and (b) are linearly independent because the Wronski matrix $\mathbf{W}(0)$ is nonsingular in each case. (c) is dependent because $\sin^2 x - \cos^2 x + \cos 2x = 0$.

4.3.14. If $\mathcal{S}$ were dependent, then there would exist a constant α such that $x^3 = \alpha|x|^3$ for all values of x. But this would mean that

$$\alpha = \frac{x^3}{|x|^3} = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \end{cases}$$

which is clearly impossible since α must be constant. The associated Wronski matrix is

$$\mathbf{W}(x) = \begin{cases} \begin{pmatrix} x^3 & x^3 \\ 3x^2 & 3x^2 \end{pmatrix} & \text{when } x \geq 0, \\ \begin{pmatrix} x^3 & -x^3 \\ 3x^2 & -3x^2 \end{pmatrix} & \text{when } x < 0, \end{cases}$$

which is singular for all values of x.

4.3.15. Start with the fact that

$$\begin{aligned} \mathbf{A}^T \text{ diag. dom.} \quad &\Longrightarrow \quad |b_{ii}| > |d_i| + \sum_{j \neq i} |b_{ji}| \quad \text{and} \quad |\alpha| > \sum_j |c_j| \\ &\Longrightarrow \quad \sum_{j \neq i} |b_{ji}| < |b_{ii}| - |d_i| \quad \text{and} \quad \frac{1}{|\alpha|} \sum_{j \neq i} |c_j| < 1 - \frac{|c_i|}{|\alpha|}, \end{aligned}$$

and then use the forward and backward triangle inequality to write

$$\begin{aligned} \sum_{j \neq i} |x_{ij}| &= \sum_{j \neq i} \left| b_{ji} - \frac{d_i c_j}{\alpha} \right| \leq \sum_{j \neq i} |b_{ji}| + \frac{|d_i|}{|\alpha|} \sum_{j \neq i} |c_j| \\ &< \big(|b_{ii}| - |d_i|\big) + |d_i| \left(1 - \frac{|c_i|}{|\alpha|}\right) = |b_{ii}| - \frac{|d_i|\,|c_i|}{|\alpha|} \\ &\leq \left| b_{ii} - \frac{d_i c_i}{\alpha} \right| = |x_{ii}|. \end{aligned}$$

Now, diagonal dominance of $\mathbf{A}^T$ insures that α is the entry of largest magnitude in the first column of $\mathbf{A}$, so no row interchange is needed at the first step of Gaussian elimination. After one step, the diagonal dominance of $\mathbf{X}$ guarantees that the magnitude of the second pivot is maximal with respect to row interchanges. Proceeding by induction establishes that no step requires partial pivoting.

Solutions for exercises in section 4. 4

4.4.1. $\dim R(\mathbf{A}) = \dim R\left(\mathbf{A}^T\right) = rank(\mathbf{A}) = 2, \quad \dim N(\mathbf{A}) = n - r = 4 - 2 = 2,$ and $\dim N\left(\mathbf{A}^T\right) = m - r = 3 - 2 = 1.$

4.4.2. $\mathcal{B}_{R(\mathbf{A})} = \left\{ \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{B}_{N(\mathbf{A}^T)} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \right\}$

$$\mathcal{B}_{R(\mathbf{A}^T)} = \left\{ \begin{pmatrix} 1 \\ 2 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 3 \\ 3 \end{pmatrix} \right\}, \quad \mathcal{B}_{N(\mathbf{A})} = \left\{ \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -2 \\ 0 \\ -3 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ -3 \\ 0 \\ 1 \end{pmatrix} \right\}$$

4.4.3. $\dim\left(span(\mathcal{S})\right) = 3$

4.4.4. (a) $n+1$ (See Example 4.4.1) (b) mn (c) $\frac{n^2+n}{2}$

4.4.5. Use the technique of Example 4.4.5. Find $\mathbf{E_A}$ to determine

$$\left\{ \mathbf{h}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{h}_2 = \begin{pmatrix} -2 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \mathbf{h}_3 = \begin{pmatrix} -1 \\ 0 \\ -2 \\ 0 \\ 1 \end{pmatrix} \right\}$$

is a basis for $N(\mathbf{A})$. Reducing the matrix $\left(\mathbf{v},\ \mathbf{h}_1,\ \mathbf{h}_2,\ \mathbf{h}_3\right)$ to row echelon form reveals that its first, second, and fourth columns are basic, and hence $\{\mathbf{v},\ \mathbf{h}_1,\ \mathbf{h}_3\}$ is a basis for $N(\mathbf{A})$ that contains $\mathbf{v}$.

4.4.6. Placing the vectors from $\mathcal{A}$ and $\mathcal{B}$ as rows in matrices and reducing

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 5 & 8 & 7 \\ 3 & 4 & 1 \end{pmatrix} \longrightarrow \mathbf{E_A} = \begin{pmatrix} 1 & 0 & -5 \\ 0 & 1 & 4 \\ 0 & 0 & 0 \end{pmatrix}$$

and

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

shows $\mathbf{A}$ and $\mathbf{B}$ have the same row space (recall Example 4.2.2), and hence $\mathcal{A}$ and $\mathcal{B}$ span the same space. Because $\mathcal{B}$ is linearly independent, it follows that $\mathcal{B}$ is a basis for $span(\mathcal{A})$.

4.4.7. $3 = \dim N(\mathbf{A}) = n - r = 4 - rank(\mathbf{A}) \implies rank(\mathbf{A}) = 1.$ Therefore, any rank-one matrix with no zero entries will do the job.

4.4.8. If $\mathbf{v} = \alpha_1\mathbf{b}_1 + \alpha_2\mathbf{b}_2 + \cdots + \alpha_n\mathbf{b}_n$ and $\mathbf{v} = \beta_1\mathbf{b}_1 + \beta_2\mathbf{b}_2 + \cdots + \beta_n\mathbf{b}_n,$ then subtraction produces

$$\mathbf{0} = (\alpha_1 - \beta_1)\mathbf{b}_1 + (\alpha_2 - \beta_2)\mathbf{b}_2 + \cdots + (\alpha_n - \beta_n)\mathbf{b}_n.$$

But $\mathcal{B}$ is a linearly independent set, so this equality can hold only if $(\alpha_i - \beta_i) = 0$ for each $i = 1, 2, \ldots, n$, and hence the α_i 's are unique.

4.4.9. Prove that if $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ is a basis for $\mathcal{S}$, then $\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\}$ is a basis for $\mathbf{A}(\mathcal{S})$. The result of Exercise 4.1.9 insures that

$$span\,\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\} = \mathbf{A}(\mathcal{S}),$$

so we need only establish the independence of $\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\}$. To do this, write

$$\begin{aligned}
\sum_{i=1}^{k} \alpha_i\,(\mathbf{A}\mathbf{s}_i) = \mathbf{0} \quad &\Longrightarrow \quad \mathbf{A}\left(\sum_{i=1}^{k} \alpha_i \mathbf{s}_i\right) = \mathbf{0} \quad \Longrightarrow \quad \sum_{i=1}^{k} \alpha_i \mathbf{s}_i \in N\,(\mathbf{A}) \\
&\Longrightarrow \quad \sum_{i=1}^{k} \alpha_i \mathbf{s}_i = \mathbf{0} \quad \text{because} \quad \mathcal{S} \cap N\,(\mathbf{A}) = \mathbf{0} \\
&\Longrightarrow \quad \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0
\end{aligned}$$

because $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k\}$ is linearly independent. Since $\{\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \ldots, \mathbf{A}\mathbf{s}_k\}$ is a basis for $\mathbf{A}(\mathcal{S})$, it follows that $\dim \mathbf{A}(\mathcal{S}) = k = \dim(\mathcal{S})$.

4.4.10. $rank\,(\mathbf{A}) = rank\,(\mathbf{A} - \mathbf{B} + \mathbf{B}) \leq rank\,(\mathbf{A} - \mathbf{B}) + rank\,(\mathbf{B})$ implies that

$$rank\,(\mathbf{A}) - rank\,(\mathbf{B}) \leq rank\,(\mathbf{A} - \mathbf{B}).$$

Furthermore, $rank\,(\mathbf{B}) = rank\,(\mathbf{B} - \mathbf{A} + \mathbf{A}) \leq rank\,(\mathbf{B} - \mathbf{A}) + rank\,(\mathbf{A}) = rank\,(\mathbf{A} - \mathbf{B}) + rank\,(\mathbf{A})$ implies that

$$-\Big(rank\,(\mathbf{A}) - rank\,(\mathbf{B})\Big) \leq rank\,(\mathbf{A} - \mathbf{B}).$$

4.4.11. Example 4.4.8 guarantees that $rank\,(\mathbf{A} + \mathbf{E}) \leq rank\,(\mathbf{A}) + rank\,(\mathbf{E}) = r + k$. Use Exercise 4.4.10 to write

$$rank\,(\mathbf{A} + \mathbf{E}) = rank\,(\mathbf{A} - (-\mathbf{E})) \geq rank\,(\mathbf{A}) - rank\,(-\mathbf{E}) = r - k.$$

4.4.12. Let $\mathbf{v}_1 \in \mathcal{V}$ such that $\mathbf{v}_1 \neq \mathbf{0}$. If $span\,\{\mathbf{v}_1\} = \mathcal{V}$, then $\mathcal{S}_1 = \{\mathbf{v}_1\}$ is an independent spanning set for $\mathcal{V}$, and we are finished. If $span\,\{\mathbf{v}_1\} \neq \mathcal{V}$, then there is a vector $\mathbf{v}_2 \in \mathcal{V}$ such that $\mathbf{v}_2 \notin span\,\{\mathbf{v}_1\}$, and hence the extension set $\mathcal{S}_2 = \{\mathbf{v}_1, \mathbf{v}_2\}$ is independent. If $span\,(\mathcal{S}_2) = \mathcal{V}$, then we are finished. Otherwise, we can proceed as described in Example 4.4.5 and continue to build independent extension sets $\mathcal{S}_3, \mathcal{S}_4, \ldots$. Statement (4.3.16) guarantees that the process must eventually yield a linearly independent spanning set $\mathcal{S}_k$ with $k \leq n$.

4.4.13. Since $\mathbf{0} = \mathbf{e}^T\mathbf{E} = \mathbf{E}_{1*} + \mathbf{E}_{2*} + \cdots + \mathbf{E}_{m*}$, any row can be written as a combination of the other $m - 1$ rows, so any set of $m - 1$ rows from $\mathbf{E}$ spans $N\left(\mathbf{E}^T\right)$. Furthermore, $rank\,(\mathbf{E}) = m - 1$ insures that no fewer than $m - 1$ vectors

can span $N\left(\mathbf{E}^T\right)$, and therefore any set of $m-1$ rows from $\mathbf{E}$ is a minimal spanning set, and hence a basis.

4.4.14. $\left[\mathbf{E}\mathbf{E}^T\right]_{ij} = \mathbf{E}_{i*}\left(\mathbf{E}^T\right)_{*j} = \mathbf{E}_{i*}\left(\mathbf{E}_{j*}\right)^T = \sum_k e_{ik}e_{jk}$. Observe that edge E_k touches node N_i if and only if $e_{ik} = \pm 1$ or, equivalently, $e_{ik}^2 = 1$. Thus $\left[\mathbf{E}\mathbf{E}^T\right]_{ii} = \sum_k e_{ik}^2 =$ the number of edges touching N_i. If $i \neq j$, then

$$e_{ik}e_{jk} = \begin{cases} -1 & \text{if } E_k \text{ is between } N_i \text{ and } N_j \\ 0 & \text{if } E_k \text{ is not between } N_i \text{ and } N_j \end{cases}$$

so that $\left[\mathbf{E}\mathbf{E}^T\right]_{ij} = \sum_k e_{ik}e_{jk} = -$ (the number of edges between N_i and N_j).

4.4.15. Apply (4.4.19) to $span\left(\mathcal{M} \cup \mathcal{N}\right) = span\left(\mathcal{M}\right) + span\left(\mathcal{N}\right)$ (see Exercise 4.1.7).

4.4.16. (a) Exercise 4.2.9 says $R\left(\mathbf{A} \mid \mathbf{B}\right) = R\left(\mathbf{A}\right) + R\left(\mathbf{B}\right)$. Since rank is the same as dimension of the range, (4.4.19) yields

$$\begin{aligned} rank\left(\mathbf{A} \mid \mathbf{B}\right) &= \dim R\left(\mathbf{A} \mid \mathbf{B}\right) = \dim\Big(R\left(\mathbf{A}\right) + R\left(\mathbf{B}\right)\Big) \\ &= \dim R\left(\mathbf{A}\right) + \dim R\left(\mathbf{B}\right) - \dim\Big(R\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)\Big) \\ &= rank\left(\mathbf{A}\right) + rank\left(\mathbf{B}\right) - \dim\Big(R\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)\Big). \end{aligned}$$

(b) Use the results of part (a) to write

$$\begin{aligned} \dim N\left(\mathbf{A} \mid \mathbf{B}\right) &= n + k - rank\left(\mathbf{A} \mid \mathbf{B}\right) \\ &= \Big(n - rank\left(\mathbf{A}\right)\Big) + \Big(k - rank\left(\mathbf{B}\right)\Big) + \dim\Big(R\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)\Big) \\ &= \dim N\left(\mathbf{A}\right) + \dim N\left(\mathbf{B}\right) + \dim\Big(R\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)\Big). \end{aligned}$$

(c) Let $\mathbf{A} = \begin{pmatrix} -1 & 1 & -2 \\ -1 & 0 & -4 \\ -1 & 0 & -5 \\ -1 & 0 & -6 \\ -1 & 0 & -6 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 3 & -2 \\ 2 & -1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$ contain bases for $R\left(\mathbf{C}\right)$ and $N\left(\mathbf{C}\right)$, respectively, so that $R\left(\mathbf{A}\right) = R\left(\mathbf{C}\right)$ and $R\left(\mathbf{B}\right) = N\left(\mathbf{C}\right)$. Use either part (a) or part (b) to obtain

$$\dim\Big(R\left(\mathbf{C}\right) \cap N\left(\mathbf{C}\right)\Big) = \dim\Big(R\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right)\Big) = 2.$$

Using $R\left(\mathbf{A} \mid \mathbf{B}\right) = R\left(\mathbf{A}\right) + R\left(\mathbf{B}\right)$ produces

$$\dim\Big(R\left(\mathbf{C}\right) + N\left(\mathbf{C}\right)\Big) = \dim\Big(R\left(\mathbf{A}\right) + R\left(\mathbf{B}\right)\Big) = rank\left(\mathbf{A} \mid \mathbf{B}\right) = 3.$$

4.4.17. Suppose $\mathbf{A}$ is $m \times n$. Existence of a solution for every $\mathbf{b}$ implies $R\left(\mathbf{A}\right) = \Re^m$. Recall from §2.5 that uniqueness of the solution implies $rank\left(\mathbf{A}\right) = n$. Thus $m = \dim R\left(\mathbf{A}\right) = rank\left(\mathbf{A}\right) = n$ so that $\mathbf{A}$ is $m \times m$ of rank m.

4.4.18. (a) $\mathbf{x} \in \mathcal{S} \implies \mathbf{x} \in span\,(\mathcal{S}_{max})$—otherwise, the extension set $\mathcal{E} = \mathcal{S}_{max} \cup \{\mathbf{x}\}$ would be linearly independent—which is impossible because $\mathcal{E}$ would contain more independent solutions than $\mathcal{S}_{max}$. Now show $span\,(\mathcal{S}_{max}) \subseteq span\,\{\mathbf{p}\} + N\,(\mathbf{A})$. Since $\mathcal{S} = \mathbf{p} + N\,(\mathbf{A})$ (see Exercise 4.2.10), $\mathbf{s}_i \in \mathcal{S}$ means there must exist a corresponding vector $\mathbf{n}_i \in N\,(\mathbf{A})$ such that $\mathbf{s}_i = \mathbf{p} + \mathbf{n}_i$, and hence

$$\begin{aligned}
\mathbf{x} \in span\,(\mathcal{S}_{max}) \quad &\implies \quad \mathbf{x} = \sum_{i=1}^{t} \alpha_i \mathbf{s}_i = \sum_{i=1}^{t} \alpha_i\,(\mathbf{p} + \mathbf{n}_i) = \sum_{i=1}^{t} \alpha_i \mathbf{p} + \sum_{i=1}^{t} \alpha_i \mathbf{n}_i \\
&\implies \quad \mathbf{x} \in span\,\{\mathbf{p}\} + N\,(\mathbf{A}) \\
&\implies \quad span\,(\mathcal{S}_{max}) \subseteq span\,\{\mathbf{p}\} + N\,(\mathbf{A}).
\end{aligned}$$

To prove the reverse inclusion, observe that if $\mathbf{x} \in span\,\{\mathbf{p}\} + N\,(\mathbf{A})$, then there exists a scalar α and a vector $\mathbf{n} \in N\,(\mathbf{A})$ such that

$$\mathbf{x} = \alpha\mathbf{p} + \mathbf{n} = (\alpha - 1)\mathbf{p} + (\mathbf{p} + \mathbf{n}).$$

Because $\mathbf{p}$ and $(\mathbf{p} + \mathbf{n})$ are both solutions, $\mathcal{S} \subseteq span(\mathcal{S}_{\max})$ guarantees that $\mathbf{p}$ and $(\mathbf{p} + \mathbf{n})$ each belong to $span\,(\mathcal{S}_{max})$, and the closure properties of a subspace insure that $\mathbf{x} \in span\,(\mathcal{S}_{max})$. Thus $span\,\{\mathbf{p}\} + N\,(\mathbf{A}) \subseteq span\,(\mathcal{S}_{max})$.
(b) The problem is really to determine the value of t in $\mathcal{S}_{max}$. The fact that $\mathcal{S}_{max}$ is a basis for $span\,(\mathcal{S}_{max})$ together with (4.4.19) produces

$$\begin{aligned}
t &= \dim\big(span\,(\mathcal{S}_{max})\big) = \dim\Big(span\,\{\mathbf{p}\} + N\,(\mathbf{A})\Big) \\
&= \dim\big(span\,\{\mathbf{p}\}\big) + \dim N\,(\mathbf{A}) - \dim\Big(span\,\{\mathbf{p}\} \cap N\,(\mathbf{A})\Big) \\
&= 1 + (n - r) - 0.
\end{aligned}$$

4.4.19. To show $\mathcal{S}_{max}$ is linearly independent, suppose

$$\mathbf{0} = \alpha_0\mathbf{p} + \sum_{i=1}^{n-r} \alpha_i\,(\mathbf{p} + \mathbf{h}_i) = \left(\sum_{i=0}^{n-r} \alpha_i\right)\mathbf{p} + \sum_{i=1}^{n-r} \alpha_i \mathbf{h}_i.$$

Multiplication by $\mathbf{A}$ yields $\mathbf{0} = \left(\sum_{i=0}^{n-r} \alpha_i\right)\mathbf{b}$, which implies $\sum_{i=0}^{n-r} \alpha_i = 0$, and hence $\sum_{i=1}^{n-r} \alpha_i \mathbf{h}_i = \mathbf{0}$. Because $\mathcal{H}$ is independent, we may conclude that $\alpha_1 = \alpha_2 = \cdots = \alpha_{n-r} = 0$. Consequently, $\alpha_0\mathbf{p} = \mathbf{0}$, and therefore $\alpha_0 = 0$ (because $\mathbf{p} \neq \mathbf{0}$), so that $\mathcal{S}_{max}$ is an independent set. By Exercise 4.4.18, it must also be maximal because it contains $n - r + 1$ vectors.

4.4.20. The proof depends on the observation that if $\mathbf{B} = \mathbf{P}^T\mathbf{A}\mathbf{P}$, where $\mathbf{P}$ is a permutation matrix, then the graph $\mathcal{G}(\mathbf{B})$ is the same as $\mathcal{G}(\mathbf{A})$ except that the nodes in $\mathcal{G}(\mathbf{B})$ have been renumbered according to the permutation defining $\mathbf{P}$. This follows because $\mathbf{P}^T = \mathbf{P}^{-1}$ implies $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^T$, so if the rows (and

columns) in $\mathbf{P}$ are the unit vectors that appear according to the permutation $\boldsymbol{\pi} = \begin{pmatrix} 1 & 2 & \cdots & n \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix}$, then

$$a_{ij} = \left[\mathbf{PBP}^T\right]_{ij} = \left[\begin{pmatrix} \mathbf{e}_{\pi_1}^T \\ \vdots \\ \mathbf{e}_{\pi_n}^T \end{pmatrix} \mathbf{B} \left(\mathbf{e}_{\pi_1} \quad \cdots \quad \mathbf{e}_{\pi_n} \right)\right]_{ij} = \mathbf{e}_{\pi_i}^T \mathbf{B} \mathbf{e}_{\pi_j} = b_{\pi_i \pi_j}.$$

Consequently, $a_{ij} \neq 0$ if and only if $b_{\pi_i \pi_j} \neq 0$, and thus $\mathcal{G}(\mathbf{A})$ and $\mathcal{G}(\mathbf{B})$ are the same except for the fact that node N_k in $\mathcal{G}(\mathbf{A})$ is node N_{π_k} in $\mathcal{G}(\mathbf{B})$ for each $k = 1, 2, \ldots, n$. Now we can prove $\mathcal{G}(\mathbf{A})$ is *not* strongly connected $\Longleftrightarrow$ $\mathbf{A}$ is reducible. If $\mathbf{A}$ is reducible, then there is a permutation matrix such that $\mathbf{P}^T\mathbf{AP} = \mathbf{B} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$, where $\mathbf{X}$ is $r \times r$ and $\mathbf{Z}$ is $n - r \times n - r$. The zero pattern in $\mathbf{B}$ indicates that the nodes $\{N_1, N_2, \ldots, N_r\}$ in $\mathcal{G}(\mathbf{B})$ are inaccessible from nodes $\{N_{r+1}, N_{r+2}, \ldots, N_n\}$, and hence $\mathcal{G}(\mathbf{B})$ is not strongly connected—e.g., there is no sequence of edges leading from N_{r+1} to N_1. Since $\mathcal{G}(\mathbf{B})$ is the same as $\mathcal{G}(\mathbf{A})$ except that the nodes have different numbers, we may conclude that $\mathcal{G}(\mathbf{A})$ is also not strongly connected. Conversely, if $\mathcal{G}(\mathbf{A})$ is not strongly connected, then there are two nodes in $\mathcal{G}(\mathbf{A})$ such that one is inaccessible from the other by any sequence of directed edges. Relabel the nodes in $\mathcal{G}(\mathbf{A})$ so that this pair is N_1 and N_n, where N_1 is inaccessible from N_n. If there are additional nodes—excluding N_n itself—which are also inaccessible from N_n, label them $N_2, N_3, \ldots, N_r$ so that the set of all nodes that are inaccessible from N_n—with the possible exception of N_n itself—is $\overline{N_n} = \{N_1, N_2, \ldots, N_r\}$ (inaccessible nodes). Label the remaining nodes—which are all accessible from N_n—as $\underline{N_n} = \{N_{r+1}, N_{r+2}, \ldots, N_{n-1}\}$ (accessible nodes). It follows that no node in $\overline{N_n}$ can be accessible from any node in $\underline{N_n}$, for otherwise nodes in $\overline{N_n}$ would be accessible from N_n through nodes in $\underline{N_n}$. In other words, if $N_{r+k} \in \underline{N_n}$ and $N_{r+k} \to N_i \in \overline{N_n}$, then $N_n \to N_{r+k} \to N_i$, which is impossible. This means that if $\boldsymbol{\pi} = \begin{pmatrix} 1 & 2 & \cdots & n \\ \pi_1 & \pi_2 & \cdots & \pi_n \end{pmatrix}$ is the permutation generated by the relabeling process, then $a_{\pi_i \pi_j} = 0$ for each $i = r+1, r+2, \ldots, n-1$ and $j = 1, 2, \ldots, r$. Therefore, if $\mathbf{B} = \mathbf{P}^T\mathbf{AP}$, where $\mathbf{P}$ is the permutation matrix corresponding to the permutation $\boldsymbol{\pi}$, then $b_{ij} = a_{\pi_i \pi_j}$, so $\mathbf{P}^T\mathbf{AP} = \mathbf{B} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$, where $\mathbf{X}$ is $r \times r$ and $\mathbf{Z}$ is $n - r \times n - r$, and thus $\mathbf{A}$ is reducible.

Solutions for exercises in section 4. 5

4.5.1. $rank\left(\mathbf{A}^T\mathbf{A}\right) = rank\left(\mathbf{A}\right) = rank\left(\mathbf{AA}^T\right) = 2$

4.5.2. $\dim N\left(\mathbf{A}\right) \cap R\left(\mathbf{B}\right) = rank\left(\mathbf{B}\right) - rank\left(\mathbf{AB}\right) = 2 - 1 = 1.$

4.5.3. Gaussian elimination yields $\mathbf{X} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \\ 2 & 2 \end{pmatrix}$, $\mathbf{V} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and $\mathbf{XV} = \begin{pmatrix} 2 \\ 0 \\ 4 \end{pmatrix}$.

4.5.4. Statement (4.5.2) says that the rank of a product cannot exceed the rank of any factor.

4.5.5. $rank\,(\mathbf{A}) = rank\,(\mathbf{A}^T\mathbf{A}) = 0 \implies \mathbf{A} = \mathbf{0}$.

4.5.6. $rank\,(\mathbf{A}) = 2$, and there are *six* 2×2 nonsingular submatrices in $\mathbf{A}$.

4.5.7. Yes. $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$ is one of many examples.

4.5.8. No—it is not difficult to construct a counterexample using two singular matrices. If either matrix is nonsingular, then the statement is true.

4.5.9. Transposition does not alter rank, so (4.5.1) says

$$\begin{aligned} rank\,(\mathbf{AB}) &= rank(\mathbf{AB})^T = rank\,(\mathbf{B}^T\mathbf{A}^T) \\ &= rank\,(\mathbf{A}^T) - \dim N\,(\mathbf{B}^T) \cap R\,(\mathbf{A}^T) \\ &= rank\,(\mathbf{A}) - \dim N\,(\mathbf{B}^T) \cap R\,(\mathbf{A}^T). \end{aligned}$$

4.5.10. This follows immediately from (4.5.1) because $\dim N\,(\mathbf{AB}) = p - rank\,(\mathbf{AB})$ and $\dim N\,(\mathbf{B}) = p - rank\,(\mathbf{B})$.

4.5.11. (a) First notice that $N\,(\mathbf{B}) \subseteq N\,(\mathbf{AB})$ (Exercise 4.2.12) for all conformable $\mathbf{A}$ and $\mathbf{B}$, so, by (4.4.5), $\dim N\,(\mathbf{B}) \leq \dim N\,(\mathbf{AB})$, or $\nu(\mathbf{B}) \leq \nu(\mathbf{AB})$, is always true—this also answers the second half of part (b). If $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$, then the rank-plus-nullity theorem together with (4.5.2) produces

$$\nu(\mathbf{A}) = \dim N\,(\mathbf{A}) = n - rank\,(\mathbf{A}) \leq n - rank\,(\mathbf{AB}) = \dim N\,(\mathbf{AB}) = \nu(\mathbf{AB}),$$

so, together with the first observation, we have $\max\{\nu(\mathbf{A}),\ \nu(\mathbf{B})\} \leq \nu(\mathbf{AB})$. The rank-plus-nullity theorem applied to (4.5.3) yields $\nu(\mathbf{AB}) \leq \nu(\mathbf{A}) + \nu(\mathbf{B})$.

(b) To see that $\nu(\mathbf{A}) > \nu(\mathbf{AB})$ is possible for rectangular matrices, consider $\mathbf{A} = (1 \quad 1)$ and $\mathbf{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

4.5.12. (a) $rank\,(\mathbf{B}_{n\times p}) = n \implies R\,(\mathbf{B}) = \Re^n \implies N\,(\mathbf{A}) \cap R\,(\mathbf{B}) = N\,(\mathbf{A}) \implies$

$$\begin{aligned} rank\,(\mathbf{AB}) &= rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) = n - \dim N\,(\mathbf{A}) \\ &= n - (n - rank\,(\mathbf{A})) = rank\,(\mathbf{A}) \end{aligned}$$

It's always true that $R\,(\mathbf{AB}) \subseteq R\,(\mathbf{A})$. When $\dim R\,(\mathbf{AB}) = \dim R\,(\mathbf{A})$ (i.e., when $rank\,(\mathbf{AB}) = rank\,(\mathbf{A})$), (4.4.6) implies $R\,(\mathbf{AB}) = R\,(\mathbf{A})$.

(b) $rank\,(\mathbf{A}_{m\times n}) = n \implies N\,(\mathbf{A}) = \{\mathbf{0}\} \implies N\,(\mathbf{A}) \cap R\,(\mathbf{B}) = \{\mathbf{0}\} \implies$

$$rank\,(\mathbf{AB}) = rank\,(\mathbf{B}) - \dim N\,(\mathbf{A}) \cap R\,(\mathbf{B}) = rank\,(\mathbf{B})$$

Assuming the product exists, it is always the case that $N\,(\mathbf{B}) \subseteq N\,(\mathbf{AB})$. Use $rank\,(\mathbf{B}) = rank\,(\mathbf{AB}) \implies p - rank\,(\mathbf{B}) = p - rank\,(\mathbf{AB}) \implies \dim N\,(\mathbf{B}) = \dim N\,(\mathbf{AB})$ together with (4.4.6) to conclude that $N\,(\mathbf{B}) = N\,(\mathbf{AB})$.

4.5.13. (a) $rank\,(\mathbf{A}) = 2$, and the unique exact solution is $(-1,\ 1)$.

(b) Same as part (a).

(c) The 3-digit rank is 2, and the unique 3-digit solution is $(-1,\ 1)$.

(d) The 3-digit normal equations $\begin{pmatrix} 6 & 12 \\ 12 & 24 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 6.01 \\ 12 \end{pmatrix}$ have infinitely many 3-digit solutions.

4.5.14. Use an indirect argument. Suppose $\mathbf{x} \in N\,(\mathbf{I}+\mathbf{F})$ in which $x_i \neq 0$ is a component of maximal magnitude. Use the triangle inequality together with $\mathbf{x} = -\mathbf{F}\mathbf{x}$ to conclude

$$|x_i| = \Big|\sum_{j=1}^{r} f_{ij}x_j\Big| \leq \sum_{j=1}^{r} |f_{ij}x_j| = \sum_{j=1}^{r} |f_{ij}|\,|x_j| \leq \Big(\sum_{j=1}^{r} |f_{ij}|\Big)|x_i| < |x_i|,$$

which is impossible. Therefore, $N\,(\mathbf{I}+\mathbf{F}) = \mathbf{0}$, and hence $\mathbf{I}+\mathbf{F}$ is nonsingular.

4.5.15. Follow the approach used in (4.5.8) to write

$$\mathbf{A} \sim \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}, \quad \text{where} \quad \mathbf{S} = \mathbf{Z} - \mathbf{Y}\mathbf{W}^{-1}\mathbf{X}.$$

$rank\,(\mathbf{A}) = rank\,(\mathbf{W}) \implies rank\,(\mathbf{S}) = 0 \implies \mathbf{S} = \mathbf{0}$, so $\mathbf{Z} = \mathbf{Y}\mathbf{W}^{-1}\mathbf{X}$. The desired conclusion now follows by taking $\mathbf{B} = \mathbf{Y}\mathbf{W}^{-1}$ and $\mathbf{C} = \mathbf{W}^{-1}\mathbf{X}$.

4.5.16. (a) Suppose that $\mathbf{A}$ is nonsingular, and let $\mathbf{E}_k = \mathbf{A}_k - \mathbf{A}$ so that $\lim_{k\to\infty} \mathbf{E}_k = \mathbf{0}$. This together with (4.5.9) implies there exists a sufficiently large value of k such that

$$rank\,(\mathbf{A}_k) = rank\,(\mathbf{A}+\mathbf{E}_k) \geq rank\,(\mathbf{A}) = n,$$

which is impossible because each $\mathbf{A}_k$ is singular. Therefore, the supposition that $\mathbf{A}$ is nonsingular must be false.

(b) No!—consider $\left[\frac{1}{k}\right]_{1\times 1} \to [0]$.

4.5.17. $\mathcal{M} \subseteq \mathcal{N}$ because $R\,(\mathbf{BC}) \subseteq R\,(\mathbf{B})$, and therefore $\dim \mathcal{M} \leq \dim \mathcal{N}$. Formula (4.5.1) guarantees $\dim \mathcal{M} = rank\,(\mathbf{BC}) - rank\,(\mathbf{ABC})$ and $\dim \mathcal{N} = rank\,(\mathbf{B}) - rank\,(\mathbf{AB})$, so the desired conclusion now follows.

4.5.18. $N\,(\mathbf{A}) \subseteq N\left(\mathbf{A}^2\right)$ and $R\left(\mathbf{A}^2\right) \subseteq R\,(\mathbf{A})$ always hold, so (4.4.6) insures

$$\begin{aligned} N\,(\mathbf{A}) = N\left(\mathbf{A}^2\right) &\iff \dim N\,(\mathbf{A}) = \dim N\left(\mathbf{A}^2\right) \\ &\iff n - rank\,(\mathbf{A}) = n - rank\left(\mathbf{A}^2\right) \\ &\iff rank\,(\mathbf{A}) = rank\left(\mathbf{A}^2\right) \\ &\iff R\,(\mathbf{A}) = R\left(\mathbf{A}^2\right). \end{aligned}$$

Formula (4.5.1) says $rank\left(\mathbf{A}^2\right) = rank\,(\mathbf{A}) - \dim R\,(\mathbf{A}) \cap N\,(\mathbf{A})$, so $R\left(\mathbf{A}^2\right) = R\,(\mathbf{A}) \iff rank\left(\mathbf{A}^2\right) = rank\,(\mathbf{A}) \iff \dim R\,(\mathbf{A}) \cap N\,(\mathbf{A}) = 0$.

4.5.19. (a) Since

$$\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{A}+\mathbf{B})(\mathbf{A} \mid \mathbf{B}) = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{A} \mid \mathbf{B}) = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix},$$

the result of Example 3.9.3 together with (4.5.2) insures

$$rank\,(\mathbf{A}) + rank\,(\mathbf{B}) \leq rank\,(\mathbf{A}+\mathbf{B}).$$

Couple this with the fact that $rank\,(\mathbf{A}+\mathbf{B}) \leq rank\,(\mathbf{A}) + rank\,(\mathbf{B})$ (see Example 4.4.8) to conclude $rank\,(\mathbf{A}+\mathbf{B}) = rank\,(\mathbf{A}) + rank\,(\mathbf{B})$.
(b) Verify that if $\mathbf{B} = \mathbf{I} - \mathbf{A}$, then $\mathbf{B}^2 = \mathbf{B}$ and $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$, and apply the result of part (a).

4.5.20. (a) $\mathbf{B}^T\mathbf{A}\mathbf{C}^T = \mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T$. The products $\mathbf{B}^T\mathbf{B}$ and $\mathbf{C}\mathbf{C}^T$ are each nonsingular because they are $r \times r$ with

$$rank\left(\mathbf{B}^T\mathbf{B}\right) = rank\,(\mathbf{B}) = r \quad \text{and} \quad rank\left(\mathbf{C}\mathbf{C}^T\right) = rank\,(\mathbf{C}) = r.$$

(b) Notice that $\mathbf{A}^\dagger = \mathbf{C}^T\left(\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T\right)^{-1}\mathbf{B}^T = \mathbf{C}^T\left(\mathbf{C}\mathbf{C}^T\right)^{-1}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T$, so

$$\mathbf{A}^T\mathbf{A}\mathbf{A}^\dagger\mathbf{b} = \mathbf{C}^T\mathbf{B}^T\mathbf{B}\mathbf{C}\mathbf{C}^T\left(\mathbf{C}\mathbf{C}^T\right)^{-1}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{b} = \mathbf{C}^T\mathbf{B}^T\mathbf{b} = \mathbf{A}^T\mathbf{b}.$$

If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent, then its solution set agrees with the solution set for the normal equations.
(c) $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{B}\mathbf{C}\mathbf{C}^T\left(\mathbf{C}\mathbf{C}^T\right)^{-1}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{B}\mathbf{C} = \mathbf{B}\mathbf{C} = \mathbf{A}$. Now,

$$\begin{aligned} \mathbf{x} \in R\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right) &\implies \mathbf{x} = \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{y} \quad \text{for some } \mathbf{y} \\ &\implies \mathbf{A}\mathbf{x} = \left(\mathbf{A} - \mathbf{A}\mathbf{A}^\dagger\mathbf{A}\right)\mathbf{y} = \mathbf{0} \implies \mathbf{x} \in N\,(\mathbf{A}). \end{aligned}$$

Conversely,

$$\mathbf{x} \in N\,(\mathbf{A}) \implies \mathbf{A}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x} \implies \mathbf{x} \in R\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right),$$

so $R\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right) = N\,(\mathbf{A})$. As $\mathbf{h}$ ranges over all of $\Re^{n\times 1}$, the expression $\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{h}$ generates $R\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right) = N\,(\mathbf{A})$. Since $\mathbf{A}^\dagger\mathbf{b}$ is a particular solution of $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$, the general solution is

$$\mathbf{x} = \mathbf{A}^\dagger\mathbf{b} + N\,(\mathbf{A}) = \mathbf{A}^\dagger\mathbf{b} + \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{h}.$$

(d) If $r = n$, then $\mathbf{B} = \mathbf{A}$ and $\mathbf{C} = \mathbf{I}_n$.
(e) If $\mathbf{A}$ is nonsingular, then so is $\mathbf{A}^T$, and

$$\mathbf{A}^\dagger = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T = \mathbf{A}^{-1}\left(\mathbf{A}^T\right)^{-1}\mathbf{A}^T = \mathbf{A}^{-1}.$$

(f) Follow along the same line as indicated in the solution to part (c) for the case $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A}$.

Solutions for exercises in section 4. 6

4.6.1. $\mathbf{A} = \begin{pmatrix} 5 \\ 7 \\ 8 \\ 10 \\ 12 \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} 11.1 \\ 15.4 \\ 17.5 \\ 22.0 \\ 26.3 \end{pmatrix}$, so $\mathbf{A}^T\mathbf{A} = 382$ and $\mathbf{A}^T\mathbf{b} = 838.9$. Thus the least squares estimate for k is $838.9/382 = 2.196$.

4.6.2. This is essentially the same problem as Exercise 4.6.1. Because it must pass through the origin, the equation of the least squares line is $y = mx$, and hence $\mathbf{A} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ and $\mathbf{b} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, so $\mathbf{A}^T\mathbf{A} = \sum_i x_i^2$ and $\mathbf{A}^T\mathbf{b} = \sum_i x_i y_i$.

4.6.3. Look for the line $p = \alpha + \beta t$ that comes closest to the data in the least squares sense. That is, find the least squares solution for the system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 7 \\ 4 \\ 3 \end{pmatrix}.$$

Set up normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ to get

$$\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 14 \\ 24 \end{pmatrix} \implies \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 26/3 \\ -2 \end{pmatrix} \implies p = (26/3) - 2t.$$

Setting $p = 0$ gives $t = 13/3$. In other words, we expect the company to begin losing money on May 1 of year five.

4.6.4. The associated linear system $\mathbf{Ax} = \mathbf{b}$ is

$$\begin{array}{ll} \text{Year 1:} & \alpha + \beta = 1 \\ \text{Year 2:} & 2\alpha = 1 \\ \text{Year 3:} & -\beta = 1 \end{array} \quad \text{or} \quad \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The least squares solution to this inconsistent system is obtained from the system of normal equations $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ that is $\begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$. The unique solution is $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$, so the least squares estimate for the increase in bread prices is

$$B = \frac{2}{3}W - \frac{1}{3}M.$$

When $W = -1$ and $M = -1$, we estimate that $B = -1/3$.

4.6.5. (a) $\alpha_0 = .02$ and $\alpha_1 = .0983$. (b) 1.986 grams.

4.6.6. Use $\ln y = \ln \alpha_0 + \alpha_1 t$ to obtain the least squares estimates $\alpha_0 = 9.73$ and $\alpha_1 = .507$.

4.6.7. The least squares line is $y = 9.64 + .182x$ and for $\varepsilon_i = 9.64 + .182x_i - y_i$, the sum of the squares of these errors is $\sum_i \varepsilon_i^2 = 162.9$. The least squares quadratic is $y = 13.97 + .1818x - .4336x^2$, and the corresponding sum of squares of the errors is $\sum \varepsilon_i^2 = 1.622$. Therefore, we conclude that the quadratic provides a much better fit.

4.6.8. 230.7 min. $(\alpha_0 = 492.04, \alpha_1 = -23.435, \alpha_2 = -.076134, \alpha_3 = 1.8624)$

4.6.9. $\mathbf{x}_2$ is a least squares solution $\implies \mathbf{A}^T\mathbf{A}\mathbf{x}_2 = \mathbf{A}^T\mathbf{b} \implies \mathbf{0} = \mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x}_2)$. If we set $\mathbf{x}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_2$, then

$$\begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{b} - \mathbf{A}\mathbf{x}_2 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}.$$

The converse is true because

$$\begin{aligned} \begin{pmatrix} \mathbf{I}_{m\times m} & \mathbf{A} \\ \mathbf{A}^T & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} &\implies \mathbf{A}\mathbf{x}_2 = \mathbf{b} - \mathbf{x}_1 \quad \text{and} \quad \mathbf{A}^T\mathbf{x}_1 = \mathbf{0} \\ &\implies \mathbf{A}^T\mathbf{A}\mathbf{x}_2 = \mathbf{A}^T\mathbf{b} - \mathbf{A}^T\mathbf{x}_1 = \mathbf{A}^T\mathbf{b}. \end{aligned}$$

4.6.10. $\mathbf{t} \in R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^T\mathbf{A}\right) \implies \mathbf{t}^T = \mathbf{z}^T\mathbf{A}^T\mathbf{A}$ for some $\mathbf{z}$. For each $\mathbf{x}$ satisfying $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$, write

$$\hat{y} = \mathbf{t}^T\mathbf{x} = \mathbf{z}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{z}^T\mathbf{A}^T\mathbf{b},$$

and notice that $\mathbf{z}^T\mathbf{A}^T\mathbf{b}$ is independent of $\mathbf{x}$.

Solutions for exercises in section 4. 7

4.7.1. (b) and (f)

4.7.2. (a), (c), and (d)

4.7.3. Use any $\mathbf{x}$ to write $\mathbf{T}(\mathbf{0}) = \mathbf{T}(\mathbf{x} - \mathbf{x}) = \mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{x}) = \mathbf{0}$.

4.7.4. (a)

4.7.5. (a) No (b) Yes

4.7.6. $\mathbf{T}(\mathbf{u}_1) = (2, 2) = 2\mathbf{u}_1 + 0\mathbf{u}_2$ and $\mathbf{T}(\mathbf{u}_2) = (3, 6) = 0\mathbf{u}_1 + 3\mathbf{u}_2$ so that $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$.

4.7.7. (a) $[\mathbf{T}]_{\mathcal{S}\mathcal{S}'} = \begin{pmatrix} 1 & 3 \\ 0 & 0 \\ 2 & -4 \end{pmatrix}$ (b) $[\mathbf{T}]_{\mathcal{S}\mathcal{S}''} = \begin{pmatrix} 2 & -4 \\ 0 & 0 \\ 1 & 3 \end{pmatrix}$

4.7.8. $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 1 & -3/2 & 1/2 \\ -1 & 1/2 & 1/2 \\ 0 & 1/2 & -1/2 \end{pmatrix}$ and $[\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$.

4.7.9. According to (4.7.4), the jth column of $[\mathbf{T}]_{\mathcal{S}}$ is

$$[\mathbf{T}(\mathbf{e}_j)]_{\mathcal{S}} = [\mathbf{A}\mathbf{e}_j]_{\mathcal{S}} = [\mathbf{A}_{*j}]_{\mathcal{S}} = \mathbf{A}_{*j}.$$

4.7.10. $[\mathbf{T}^k]_{\mathcal{B}} = [\mathbf{T}\mathbf{T}\cdots\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}[\mathbf{T}]_{\mathcal{B}}\cdots[\mathbf{T}]_{\mathcal{B}} = [\mathbf{T}]_{\mathcal{B}}^k$

4.7.11. (a) Sketch a picture to observe that $\mathbf{P}(\mathbf{e}_1) = \begin{pmatrix} x \\ x \end{pmatrix} = \mathbf{P}(\mathbf{e}_2)$ and that the vectors $\mathbf{e}_1$, $\mathbf{P}(\mathbf{e}_1)$, and $\mathbf{0}$ are vertices of a 45° right triangle (as are $\mathbf{e}_2$, $\mathbf{P}(\mathbf{v}_2)$, and $\mathbf{0}$). So, if $\|\star\|$ denotes length, the Pythagorean theorem may be applied to yield $1 = 2\|\mathbf{P}(\mathbf{e}_1)\|^2 = 4x^2$ and $1 = 2\|\mathbf{P}(\mathbf{e}_2)\|^2 = 4x^2$. Thus

$$\left\{\begin{array}{l} \mathbf{P}(\mathbf{e}_1) = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = (1/2)\mathbf{e}_1 + (1/2)\mathbf{e}_2 \\ \mathbf{P}(\mathbf{e}_2) = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} = (1/2)\mathbf{e}_1 + (1/2)\mathbf{e}_2 \end{array}\right\} \implies [\mathbf{P}]_{\mathcal{S}} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

(b) $\mathbf{P}(\mathbf{v}) = \begin{pmatrix} \frac{\alpha+\beta}{2} \\ \frac{\alpha+\beta}{2} \end{pmatrix}$

4.7.12. (a) If $\mathbf{U}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\mathbf{U}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $\mathbf{U}_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, $\mathbf{U}_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, then

$$\begin{aligned} \mathbf{T}(\mathbf{U}_1) &= \mathbf{U}_1 + 0\mathbf{U}_2 + 0\mathbf{U}_3 + 0\mathbf{U}_4, \\ \mathbf{T}(\mathbf{U}_2) &= \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 0\mathbf{U}_1 + 1/2\mathbf{U}_2 + 1/2\mathbf{U}_3 + 0\mathbf{U}_4, \\ \mathbf{T}(\mathbf{U}_3) &= \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 0\mathbf{U}_1 + 1/2\mathbf{U}_2 + 1/2\mathbf{U}_3 + 0\mathbf{U}_4, \\ \mathbf{T}(\mathbf{U}_4) &= 0\mathbf{U}_1 + 0\mathbf{U}_2 + 0\mathbf{U}_3 + \mathbf{U}_4, \end{aligned}$$

so $[\mathbf{T}]_{\mathcal{S}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. To verify $[\mathbf{T}(\mathbf{U})]_{\mathcal{S}} = [\mathbf{T}]_{\mathcal{S}}[\mathbf{U}]_{\mathcal{S}}$, observe that

$$\mathbf{T}(\mathbf{U}) = \begin{pmatrix} a & (b+c)/2 \\ (b+c)/2 & d \end{pmatrix}, \quad [\mathbf{T}(\mathbf{U})]_{\mathcal{S}} = \begin{pmatrix} a \\ (b+c)/2 \\ (b+c)/2 \\ d \end{pmatrix}, \quad [\mathbf{U}]_{\mathcal{S}} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}.$$

(b) For $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, and $\mathbf{U}_4$ as defined above,

$$\mathbf{T}(\mathbf{U}_1) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} = 0\mathbf{U}_1 - \mathbf{U}_2 - \mathbf{U}_3 + 0\mathbf{U}_4,$$

$$\mathbf{T}(\mathbf{U}_2) = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} = \mathbf{U}_1 + 2\mathbf{U}_2 + 0\mathbf{U}_3 - \mathbf{U}_4,$$

$$\mathbf{T}(\mathbf{U}_3) = \begin{pmatrix} 1 & 0 \\ -2 & -1 \end{pmatrix} = \mathbf{U}_1 + 0\mathbf{U}_2 - 2\mathbf{U}_3 - 1\mathbf{U}_4,$$

$$\mathbf{T}(\mathbf{U}_4) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 0\mathbf{U}_1 + \mathbf{U}_2 + \mathbf{U}_3 + 0\mathbf{U}_4,$$

so $[\mathbf{T}]_{\mathcal{S}} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ -1 & 2 & 0 & 1 \\ -1 & 0 & -2 & 1 \\ 0 & -1 & -1 & 0 \end{pmatrix}$. To verify $[\mathbf{T}(\mathbf{U})]_{\mathcal{S}} = [\mathbf{T}]_{\mathcal{S}}[\mathbf{U}]_{\mathcal{S}}$, observe that

$$\mathbf{T}(\mathbf{U}) = \begin{pmatrix} c+b & -a+2b+d \\ -a-2c+d & -b-c \end{pmatrix} \text{ and } [\mathbf{T}(\mathbf{U})]_{\mathcal{S}} = \begin{pmatrix} c+b \\ -a+2b+d \\ -a-2c+d \\ -b-c \end{pmatrix}.$$

4.7.13. $[\mathbf{S}]_{\mathcal{B}\mathcal{B}'} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}$

4.7.14. (a) $[\mathbf{RQ}]_{\mathcal{S}} = [\mathbf{R}]_{\mathcal{S}}[\mathbf{Q}]_{\mathcal{S}} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ -\sin\theta & -\cos\theta \end{pmatrix}$

(b)

$$[\mathbf{QQ}]_{\mathcal{S}} = [\mathbf{Q}]_{\mathcal{S}}[\mathbf{Q}]_{\mathcal{S}} = \begin{pmatrix} \cos^2\theta - \sin^2\theta & -2\cos\theta\sin\theta \\ 2\cos\theta\sin\theta & \cos^2\theta - \sin^2\theta \end{pmatrix}$$
$$= \begin{pmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{pmatrix}$$

4.7.15. (a) Let $\mathcal{B} = \{\mathbf{u}_i\}_{i=1}^n$, $\mathcal{B}' = \{\mathbf{v}_i\}_{i=1}^m$. If $[\mathbf{P}]_{\mathcal{B}\mathcal{B}'} = [\alpha_{ij}]$ and $[\mathbf{Q}]_{\mathcal{B}\mathcal{B}'} = [\beta_{ij}]$, then $\mathbf{P}(\mathbf{u}_j) = \sum_i \alpha_{ij}\mathbf{v}_i$ and $\mathbf{Q}(\mathbf{u}_j) = \sum_i \beta_{ij}\mathbf{v}_i$. Thus $(\mathbf{P}+\mathbf{Q})(\mathbf{u}_j) = \sum_i(\alpha_{ij} + \beta_{ij})\mathbf{v}_i$ and hence $[\mathbf{P}+\mathbf{Q}]_{\mathcal{B}\mathcal{B}'} = [\alpha_{ij}+\beta_{ij}] = [\alpha_{ij}] + [\beta_{ij}] = [\mathbf{P}]_{\mathcal{B}\mathcal{B}'} + [\mathbf{Q}]_{\mathcal{B}\mathcal{B}'}$. The proof of part (b) is similar.

4.7.16. (a) If $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^n$ is a basis, then $\mathbf{I}(\mathbf{x}_j) = 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 1\mathbf{x}_j + \cdots + 0\mathbf{x}_n$ so that the jth column in $[\mathbf{I}]_{\mathcal{B}}$ is just the jth unit column.

(b) Suppose $\mathbf{x}_j = \sum_i \beta_{ij}\mathbf{y}_i$ so that $[\mathbf{x}_j]_{\mathcal{B}'} = \begin{pmatrix} \beta_{1j} \\ \vdots \\ \beta_{nj} \end{pmatrix}$. Then

$$\mathbf{I}(\mathbf{x}_j) = \mathbf{x}_j = \sum_i \beta_{ij}\mathbf{y}_i \implies [\mathbf{I}]_{\mathcal{B}\mathcal{B}'} = [\beta_{ij}] = \Big([\mathbf{x}_1]_{\mathcal{B}'} \,\Big|\, [\mathbf{x}_2]_{\mathcal{B}'} \,\Big|\, \cdots \,\Big|\, [\mathbf{x}_n]_{\mathcal{B}'} \Big).$$

Furthermore, $\mathbf{T}(\mathbf{y}_j) = \mathbf{x}_j = \sum_i \beta_{ij}\mathbf{y}_i \implies [\mathbf{T}]_{\mathcal{B}'} = [\beta_{ij}]$, and

$$\mathbf{T}(\mathbf{x}_j) = \mathbf{T}\Big(\sum_i \beta_{ij}\mathbf{y}_i\Big) = \sum_i \beta_{ij}\mathbf{T}(\mathbf{y}_i) = \sum_i \beta_{ij}\mathbf{x}_i \implies [\mathbf{T}]_{\mathcal{B}} = [\beta_{ij}].$$

(c) $\begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$

4.7.17. (a) $\mathbf{T}^{-1}(x, y, z) = (x + y + z,\ \ x + 2y + 2z,\ \ x + 2y + 3z)$

(b) $[\mathbf{T}^{-1}]_{\mathcal{S}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} = [\mathbf{T}]_{\mathcal{S}}^{-1}$

4.7.18. (1) $\Longrightarrow$ (2): $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y}) \Longrightarrow \mathbf{T}(\mathbf{x}-\mathbf{y}) = \mathbf{0} \Longrightarrow (\mathbf{y}-\mathbf{x}) = \mathbf{T}^{-1}(\mathbf{0}) = \mathbf{0}$.

(2) $\Longrightarrow$ (3): $\mathbf{T}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{T}(\mathbf{0}) = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{0}$.

(3) $\Longrightarrow$ (4): If $\{\mathbf{u}_i\}_{i=1}^n$ is a basis for $\mathcal{V}$, show that $N(\mathbf{T}) = \{\mathbf{0}\}$ implies $\{\mathbf{T}(\mathbf{u}_i)\}_{i=1}^n$ is also a basis. Consequently, for each $\mathbf{v} \in \mathcal{V}$ there are coordinates ξ_i such that

$$\mathbf{v} = \sum_i \xi_i \mathbf{T}(\mathbf{u}_i) = \mathbf{T}\Big(\sum_i \xi_i \mathbf{u}_i\Big).$$

(4) $\Longrightarrow$ (2): For each basis vector $\mathbf{u}_i$, there is a $\mathbf{v}_i$ such that $\mathbf{T}(\mathbf{v}_i) = \mathbf{u}_i$. Show that $\{\mathbf{v}_i\}_{i=1}^n$ is also a basis. If $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$, then $\mathbf{T}(\mathbf{x} - \mathbf{y}) = \mathbf{0}$. Let $\mathbf{x} - \mathbf{y} = \sum_i \xi_i \mathbf{v}_i$ so that $\mathbf{0} = \mathbf{T}(\mathbf{x} - \mathbf{y}) = \mathbf{T}\Big(\sum_i \xi_i \mathbf{v}_i\Big) = \sum_i \xi_i \mathbf{T}(\mathbf{v}_i) = \sum_i \xi_i \mathbf{u}_i \Longrightarrow$ each $\xi_i = 0 \Longrightarrow \mathbf{x} - \mathbf{y} = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{y}$.

(4) and (2) $\Longrightarrow$ (1): For each $\mathbf{y} \in \mathcal{V}$, show there is a *unique* $\mathbf{x}$ such that $\mathbf{T}(\mathbf{x}) = \mathbf{y}$. Let $\hat{\mathbf{T}}$ be the function defined by the rule $\hat{\mathbf{T}}(\mathbf{y}) = \mathbf{x}$. Clearly, $\mathbf{T}\hat{\mathbf{T}} = \hat{\mathbf{T}}\mathbf{T} = \mathbf{I}$. To show that $\hat{\mathbf{T}}$ is a linear function, consider $\alpha\mathbf{y}_1 + \mathbf{y}_2$, and let $\mathbf{x}_1$ and $\mathbf{x}_2$ be such that $\mathbf{T}(\mathbf{x}_1) = \mathbf{y}_1$, $\mathbf{T}(\mathbf{x}_2) = \mathbf{y}_2$. Now, $\mathbf{T}(\alpha\mathbf{x}_1 + \mathbf{x}_2) = \alpha\mathbf{y}_1 + \mathbf{y}_2$ so that $\hat{\mathbf{T}}(\alpha\mathbf{y}_1 + \mathbf{y}_2) = \alpha\mathbf{x}_1 + \mathbf{x}_2$. However, $\mathbf{x}_1 = \hat{\mathbf{T}}(\mathbf{y}_1)$, $\mathbf{x}_2 = \hat{\mathbf{T}}(\mathbf{y}_2)$ so that $\alpha\hat{\mathbf{T}}(\mathbf{y}_1) + \hat{\mathbf{T}}(\mathbf{y}_2) = \alpha\mathbf{x}_1 + \mathbf{x}_2 = \hat{\mathbf{T}}(\alpha\mathbf{y}_1 + \mathbf{y}_2)$. Therefore $\hat{\mathbf{T}} = \mathbf{T}^{-1}$.

4.7.19. (a) $\mathbf{0} = \sum_i \alpha_i \mathbf{x}_i \Longleftrightarrow \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = [\mathbf{0}]_{\mathcal{B}} = \Big[\sum_i \alpha_i \mathbf{x}_i\Big]_{\mathcal{B}} = \sum_i [\alpha_i \mathbf{x}_i]_{\mathcal{B}} = \sum_i \alpha_i [\mathbf{x}_i]_{\mathcal{B}}$

(b) $\mathcal{G} = \Big\{\mathbf{T}(\mathbf{u}_1), \mathbf{T}(\mathbf{u}_2), \ldots, \mathbf{T}(\mathbf{u}_n)\Big\}$ spans $R(\mathbf{T})$. From part (a), the set

$$\Big\{\mathbf{T}(\mathbf{u}_{b_1}), \mathbf{T}(\mathbf{u}_{b_2}), \ldots, \mathbf{T}(\mathbf{u}_{b_r})\Big\}$$

is a maximal independent subset of $\mathcal{G}$ if and only if the set

$$\Big\{[\mathbf{T}(\mathbf{u}_{b_1})]_{\mathcal{B}}, [\mathbf{T}(\mathbf{u}_{b_2})]_{\mathcal{B}}, \ldots, [\mathbf{T}(\mathbf{u}_{b_r})]_{\mathcal{B}}\Big\}$$

is a maximal linearly independent subset of

$$\big\{[\mathbf{T}(\mathbf{u}_1)]_{\mathcal{B}}, [\mathbf{T}(\mathbf{u}_2)]_{\mathcal{B}}, \ldots, [\mathbf{T}(\mathbf{u}_n)]_{\mathcal{B}}\big\},$$

which are the columns of $[\mathbf{T}]_{\mathcal{B}}$.

Solutions for exercises in section 4. 8

4.8.1. Multiplication by nonsingular matrices does not change rank.

4.8.2. $\mathbf{A} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$ and $\mathbf{B} = \mathbf{P}^{-1}\mathbf{C}\mathbf{P} \implies \mathbf{A} = (\mathbf{P}\mathbf{Q})^{-1}\mathbf{C}(\mathbf{P}\mathbf{Q})$.

4.8.3. (a) $[\mathbf{A}]_{\mathcal{S}} = \begin{pmatrix} 1 & 2 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 7 \end{pmatrix}$

(b) $[\mathbf{A}]_{\mathcal{S}'} = \begin{pmatrix} 1 & 4 & 3 \\ -1 & -2 & -9 \\ 1 & 1 & 8 \end{pmatrix}$ and $\mathbf{Q} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

4.8.4. Put the vectors from $\mathcal{B}$ into a matrix $\mathbf{Q}$ and compute

$$[\mathbf{A}]_{\mathcal{B}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} -2 & -3 & -7 \\ 7 & 9 & 12 \\ -2 & -1 & 0 \end{pmatrix}.$$

4.8.5. $[\mathbf{B}]_{\mathcal{S}} = \mathbf{B}$ and $[\mathbf{B}]_{\mathcal{S}'} = \mathbf{C}$. Therefore, $\mathbf{C} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$, where $\mathbf{Q} = \begin{pmatrix} 2 & -3 \\ -1 & 2 \end{pmatrix}$ is the change of basis matrix from $\mathcal{S}'$ to $\mathcal{S}$.

4.8.6. If $\mathcal{B} = \{\mathbf{u}, \mathbf{v}\}$ is such a basis, then $\mathbf{T}(\mathbf{u}) = 2\mathbf{u}$ and $\mathbf{T}(\mathbf{v}) = 3\mathbf{v}$. For $\mathbf{u} = (u_1, u_2)$, $\mathbf{T}(\mathbf{u}) = 2\mathbf{u}$ implies

$$\begin{aligned} -7u_1 - 15u_2 &= 2u_1 \\ 6u_1 + 12u_2 &= 2u_2, \end{aligned}$$

or

$$\begin{aligned} -9u_1 - 15u_2 &= 0 \\ 6u_1 + 10u_2 &= 0, \end{aligned}$$

so $u_1 = (-5/3)u_2$ with u_2 being free. Letting $u_2 = -3$ produces $\mathbf{u} = (5, -3)$. Similarly, a solution to $\mathbf{T}(\mathbf{v}) = 3\mathbf{v}$ is $\mathbf{v} = (-3, 2)$. $[\mathbf{T}]_{\mathcal{S}} = \begin{pmatrix} -7 & -15 \\ 6 & 12 \end{pmatrix}$ and $[\mathbf{T}]_{\mathcal{B}} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. For $\mathbf{Q} = \begin{pmatrix} 5 & -3 \\ -3 & 2 \end{pmatrix}$, $[\mathbf{T}]_{\mathcal{B}} = \mathbf{Q}^{-1}[\mathbf{T}]_{\mathcal{S}}\mathbf{Q}$.

4.8.7. If $\sin\theta = 0$, the result is trivial. Assume $\sin\theta \neq 0$. Notice that with respect to the standard basis $\mathcal{S}$, $[\mathbf{P}]_{\mathcal{S}} = \mathbf{R}$. This means that if $\mathbf{R}$ and $\mathbf{D}$ are to be similar, then there must exist a basis $\mathcal{B} = \{\mathbf{u}, \mathbf{v}\}$ such that $[\mathbf{P}]_{\mathcal{B}} = \mathbf{D}$, which implies that $\mathbf{P}(\mathbf{u}) = e^{i\theta}\mathbf{u}$ and $\mathbf{P}(\mathbf{v}) = e^{-i\theta}\mathbf{v}$. For $\mathbf{u} = (u_1, u_2)$, $\mathbf{P}(\mathbf{u}) = e^{i\theta}\mathbf{u}$ implies

$$\begin{aligned} u_1\cos\theta - u_2\sin\theta &= e^{i\theta}u_1 = u_1\cos\theta + iu_1\sin\theta \\ u_1\sin\theta + u_2\cos\theta &= e^{i\theta}u_2 = u_2\cos\theta + iu_2\sin\theta, \end{aligned}$$

or

$$\begin{aligned} iu_1 + u_2 &= 0 \\ u_1 - iu_2 &= 0, \end{aligned}$$

so $u_1 = iu_2$ with u_2 being free. Letting $u_2 = 1$ produces $\mathbf{u} = (i, 1)$. Similarly, a solution to $\mathbf{P}(\mathbf{v}) = e^{-i\theta}\mathbf{v}$ is $\mathbf{v} = (1, i)$. Now, $[\mathbf{P}]_{\mathcal{S}} = \mathbf{R}$ and $[\mathbf{P}]_{\mathcal{B}} = \mathbf{D}$ so that $\mathbf{R}$ and $\mathbf{D}$ must be similar. The coordinate change matrix from $\mathcal{B}$ to $\mathcal{S}$ is $\mathbf{Q} = \begin{pmatrix} i & 1 \\ 1 & i \end{pmatrix}$, and therefore $\mathbf{D} = \mathbf{Q}^{-1}\mathbf{R}\mathbf{Q}$.

4.8.8. (a) $\mathbf{B} = \mathbf{Q}^{-1}\mathbf{C}\mathbf{Q} \implies (\mathbf{B} - \lambda\mathbf{I}) = \mathbf{Q}^{-1}\mathbf{C}\mathbf{Q} - \lambda\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{Q}^{-1}(\mathbf{C} - \lambda\mathbf{I})\mathbf{Q}$. The result follows because multiplication by nonsingular matrices does not change rank.

(b) $\mathbf{B} = \mathbf{P}^{-1}\mathbf{D}\mathbf{P} \implies \mathbf{B} - \lambda_i\mathbf{I} = \mathbf{P}^{-1}(\mathbf{D} - \lambda_i\mathbf{I})\mathbf{P}$ and $(\mathbf{D} - \lambda_i\mathbf{I})$ is singular for each λ_i. Now use part (a).

4.8.9. $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \implies \mathbf{B}^{\mathbf{k}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\cdots\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}\mathbf{A}\cdots\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}^{\mathbf{k}}\mathbf{P}$

4.8.10. (a) $\mathbf{Y}^T\mathbf{Y}$ is nonsingular because $rank\left(\mathbf{Y}^T\mathbf{Y}\right)_{n\times n} = rank\left(\mathbf{Y}\right) = n$. If

$$[\mathbf{v}]_{\mathcal{B}} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \quad \text{and} \quad [\mathbf{v}]_{\mathcal{B}'} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix},$$

then

$$\begin{aligned} \mathbf{v} = \sum_i \alpha_i\mathbf{x}_i = \mathbf{X}[\mathbf{v}]_{\mathcal{B}} \quad &\text{and} \quad \mathbf{v} = \sum_i \beta_i\mathbf{y}_i = \mathbf{Y}[\mathbf{v}]_{\mathcal{B}'} \\ \implies \mathbf{X}[\mathbf{v}]_{\mathcal{B}} = \mathbf{Y}[\mathbf{v}]_{\mathcal{B}'} &\implies \mathbf{Y}^T\mathbf{X}[\mathbf{v}]_{\mathcal{B}} = \mathbf{Y}^T\mathbf{Y}[\mathbf{v}]_{\mathcal{B}'} \\ \implies (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{X}[\mathbf{v}]_{\mathcal{B}} &= [\mathbf{v}]_{\mathcal{B}'}. \end{aligned}$$

(b) When $m = n$, $\mathbf{Y}$ is square and $(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T = \mathbf{Y}^{-1}$ so that $\mathbf{P} = \mathbf{Y}^{-1}\mathbf{X}$.

4.8.11. (a) Because $\mathcal{B}$ contains n vectors, you need only show that $\mathcal{B}$ is linearly independent. To do this, suppose $\sum_{i=0}^{n-1}\alpha_i\mathbf{N}^i(\mathbf{y}) = \mathbf{0}$ and apply $\mathbf{N}^{n-1}$ to both sides to get $\alpha_0\mathbf{N}^{n-1}(\mathbf{y}) = \mathbf{0} \implies \alpha_0 = 0$. Now $\sum_{i=1}^{n-1}\alpha_i\mathbf{N}^i(\mathbf{y}) = \mathbf{0}$. Apply $\mathbf{N}^{n-2}$ to both sides of this to conclude that $\alpha_1 = 0$. Continue this process until you have $\alpha_0 = \alpha_1 = \cdots = \alpha_{n-1} = 0$.

(b) Any $n \times n$ nilpotent matrix of index n can be viewed as a nilpotent operator of index n on $\Re^n$. Furthermore, $\mathbf{A} = [\mathbf{A}]_{\mathcal{S}}$ and $\mathbf{B} = [\mathbf{B}]_{\mathcal{S}}$, where $\mathcal{S}$ is the standard basis. According to part (a), there are bases $\mathcal{B}$ and $\mathcal{B}'$ such that $[\mathbf{A}]_{\mathcal{B}} = \mathbf{J}$ and $[\mathbf{B}]_{\mathcal{B}'} = \mathbf{J}$. Since $[\mathbf{A}]_{\mathcal{S}} \simeq [\mathbf{A}]_{\mathcal{B}}$, it follows that $\mathbf{A} \simeq \mathbf{J}$. Similarly $\mathbf{B} \simeq \mathbf{J}$, and hence $\mathbf{A} \simeq \mathbf{B}$ by Exercise 4.8.2.

(c) Trace and rank are similarity invariants, and part (a) implies that every $n \times n$ nilpotent matrix of index n is similar to $\mathbf{J}$, and $trace\left(\mathbf{J}\right) = 0$ and $rank\left(\mathbf{J}\right) = n - 1$.

4.8.12. (a) $\mathbf{x}_i \in R\left(\mathbf{E}\right) \implies \mathbf{x}_i = \mathbf{E}(\mathbf{v}_i)$ for some $\mathbf{v}_i \implies \mathbf{E}(\mathbf{x}_i) = \mathbf{E}^2(\mathbf{v}_i) = \mathbf{E}(\mathbf{v}_i) = \mathbf{x}_i$. Since $\mathcal{B}$ contains n vectors, you need only show that $\mathcal{B}$ is linearly independent. $\mathbf{0} = \sum_i \alpha_i\mathbf{x}_i + \beta_i\mathbf{y}_i \implies \mathbf{0} = \mathbf{E}(\mathbf{0}) = \sum_i \alpha_i\mathbf{E}(\mathbf{x}_i) + \beta_i\mathbf{E}(\mathbf{y}_i) = \sum_i \alpha_i\mathbf{x}_i \implies \alpha_i\text{'s} = 0 \implies \sum_i \beta_i\mathbf{y}_i = 0 \implies \beta_i\text{'s} = 0$.

(b) Let $\mathcal{B} = \mathcal{X} \cup \mathcal{Y} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$. For $j = 1, 2, \ldots, r$, the jth column of $[\mathbf{E}]_{\mathcal{B}}$ is $[\mathbf{E}(\mathbf{b}_j)]_{\mathcal{B}} = [\mathbf{E}(\mathbf{x}_j)]_{\mathcal{B}} = \mathbf{e}_j$. For $j = r+1, r+2, \ldots, n$, $[\mathbf{E}(\mathbf{b}_j)]_{\mathcal{B}} = [\mathbf{E}(\mathbf{y}_{j-r})]_{\mathcal{B}} = [\mathbf{0}]_{\mathcal{B}} = 0$.

(c) Suppose that $\mathbf{B}$ and $\mathbf{C}$ are two idempotent matrices of rank r. If you regard them as linear operators on $\Re^n$, then, with respect to the standard basis, $[\mathbf{B}]_{\mathcal{S}} = \mathbf{B}$ and $[\mathbf{C}]_{\mathcal{S}} = \mathbf{C}$. You know from part (b) that there are bases $\mathcal{U}$ and $\mathcal{V}$ such that $[\mathbf{B}]_{\mathcal{U}} = [\mathbf{C}]_{\mathcal{V}} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{P}$. This implies that $\mathbf{B} \simeq \mathbf{P}$, and $\mathbf{P} \simeq \mathbf{C}$. From Exercise 4.8.2, it follows that $\mathbf{B} \simeq \mathbf{C}$.

(d) It follows from part (c) that $\mathbf{F} \simeq \mathbf{P} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$. Since trace and rank are similarity invariants, $trace\,(\mathbf{F}) = trace\,(\mathbf{P}) = r = rank\,(\mathbf{P}) = rank\,(\mathbf{F})$.

Solutions for exercises in section 4. 9

4.9.1. (a) Yes, because $\mathbf{T}(\mathbf{0}) = \mathbf{0}$. (b) Yes, because $\mathbf{x} \in \mathcal{V} \implies \mathbf{T}(\mathbf{x}) \in \mathcal{V}$.

4.9.2. Every subspace of $\mathcal{V}$ is invariant under $\mathbf{I}$.

4.9.3. (a) $\mathcal{X}$ is invariant because $\mathbf{x} \in \mathcal{X} \iff \mathbf{x} = (\alpha, \beta, 0, 0)$ for $\alpha, \beta \in \Re$, so

$$\mathbf{T}(\mathbf{x}) = \mathbf{T}(\alpha, \beta, 0, 0) = (\alpha + \beta,\ \beta, 0, 0) \in \mathcal{X}.$$

(b) $$\left[\mathbf{T}/_{\mathcal{X}}\right]_{\{\mathbf{e}_1, \mathbf{e}_2\}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

(c) $$[\mathbf{T}]_{\mathcal{B}} = \left(\begin{array}{cc|cc} 1 & 1 & * & * \\ 0 & 1 & * & * \\ \hline 0 & 0 & * & * \\ 0 & 0 & * & * \end{array}\right)$$

4.9.4. (a) $\mathbf{Q}$ is nonsingular. (b) $\mathcal{X}$ is invariant because

$$\mathbf{T}(\alpha_1 \mathbf{Q}_{*1} + \alpha_2 \mathbf{Q}_{*2}) = \alpha_1 \begin{pmatrix} 1 \\ 1 \\ -2 \\ 3 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 2 \\ -2 \\ 2 \end{pmatrix} = \alpha_1 \mathbf{Q}_{*1} + \alpha_2 (\mathbf{Q}_{*1} + \mathbf{Q}_{*2})$$
$$= (\alpha_1 + \alpha_2)\mathbf{Q}_{*1} + \alpha_2 \mathbf{Q}_{*2} \in span\,\{\mathbf{Q}_{*1},\ \mathbf{Q}_{*2}\}.$$

$\mathcal{Y}$ is invariant because

$$\mathbf{T}(\alpha_3 \mathbf{Q}_{*3} + \alpha_4 \mathbf{Q}_{*4}) = \alpha_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_4 \begin{pmatrix} 0 \\ 3 \\ 1 \\ -4 \end{pmatrix} = \alpha_4 \mathbf{Q}_{*3} \in span\,\{\mathbf{Q}_{*3},\ \mathbf{Q}_{*4}\}.$$

(c) According to (4.9.10), $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q}$ should be block diagonal.

(d) $\mathbf{Q}^{-1}\mathbf{T}\mathbf{Q} = \left(\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) = \begin{pmatrix} \left[\mathbf{T}/\mathcal{X}\right]_{\{\mathbf{Q}_{*1},\mathbf{Q}_{*2}\}} & \mathbf{0} \\ \mathbf{0} & \left[\mathbf{T}/\mathcal{Y}\right]_{\{\mathbf{Q}_{*3},\mathbf{Q}_{*4}\}} \end{pmatrix}$

4.9.5. If $\mathbf{A} = [\alpha_{ij}]$ and $\mathbf{C} = [\gamma_{ij}]$, then

$$\mathbf{T}(\mathbf{u}_j) = \sum_{i=1}^{r} \alpha_{ij}\mathbf{u}_i \in \mathcal{U} \quad \text{and} \quad \mathbf{T}(\mathbf{w}_j) = \sum_{i=1}^{q} \gamma_{ij}\mathbf{w}_i \in \mathcal{W}.$$

4.9.6. If $\mathcal{S}$ is the standard basis for $\Re^{n\times 1}$, and if $\mathcal{B}$ is the basis consisting of the columns of $\mathbf{P}$, then

$$[\mathbf{T}]_{\mathcal{B}} = \mathbf{P}^{-1}[\mathbf{T}]_{\mathcal{S}}\mathbf{P} = \mathbf{P}^{-1}\mathbf{T}\mathbf{P} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}.$$

(Recall Example 4.8.3.) The desired conclusion now follows from the result of Exercise 4.9.5.

4.9.7. $\mathbf{x} \in N(\mathbf{A} - \lambda\mathbf{I}) \implies (\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0} \implies \mathbf{A}\mathbf{x} = \lambda\mathbf{x} \in N(\mathbf{A} - \lambda\mathbf{I})$

4.9.8. (a) $(\mathbf{A} - \lambda\mathbf{I})$ is singular when $\lambda = -1$ and $\lambda = 3$.

(b) There are four invariant subspaces—the trivial space $\{\mathbf{0}\}$, the entire space $\Re^2$, and the two one-dimensional spaces

$$N(\mathbf{A} + \mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right\} \quad \text{and} \quad N(\mathbf{A} - 3\mathbf{I}) = span\left\{\begin{pmatrix} 1 \\ 3 \end{pmatrix}\right\}.$$

(c) $\mathbf{Q} = \begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}$

Clearly spoken, Mr. Fogg; you explain English by Greek.
— *Benjamin Franklin* (1706–1790)

Solutions for Chapter 5

Solutions for exercises in section 5. 1

5.1.1. (a) $\|\mathbf{x}\|_1 = 9, \quad \|\mathbf{x}\|_2 = 5, \quad \|\mathbf{x}\|_\infty = 4$

(b) $\|\mathbf{x}\|_1 = 5 + 2\sqrt{2}, \quad \|\mathbf{x}\|_2 = \sqrt{21}, \quad \|\mathbf{x}\|_\infty = 4$

5.1.2. (a) $\|\mathbf{u} - \mathbf{v}\| = \sqrt{31}$ (b) $\|\mathbf{u} + \mathbf{v}\| = \sqrt{27} \leq 7 = \|\mathbf{u}\| + \|\mathbf{v}\|$

(c) $|\mathbf{u}^T\mathbf{v}| = 1 \leq 10 = \|\mathbf{u}\|\ \|\mathbf{v}\|$

5.1.3. Use the CBS inequality with $\mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$.

5.1.4. (a) $\left\{\mathbf{x} \in \Re^n \ \middle|\ \|\mathbf{x}\|_2 \leq 1\right\}$ (b) $\left\{\mathbf{x} \in \Re^n \ \middle|\ \|\mathbf{x} - \mathbf{c}\|_2 \leq \rho\right\}$

5.1.5. $\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2 \implies -2\mathbf{x}^T\mathbf{y} = 2\mathbf{x}^T\mathbf{y} \implies \mathbf{x}^T\mathbf{y} = 0.$

5.1.6. $\|\mathbf{x} - \mathbf{y}\| = \|(-1)(\mathbf{y} - \mathbf{x})\| = |(-1)|\ \|\mathbf{y} - \mathbf{x}\| = \|\mathbf{y} - \mathbf{x}\|$

5.1.7. $\mathbf{x} - \mathbf{y} = \sum_{i=1}^n (x_i - y_i)\mathbf{e}_i \implies \|\mathbf{x} - \mathbf{y}\| \leq \sum_{i=1}^n |x_i - y_i|\ \|\mathbf{e}_i\| \leq \nu \sum_{i=1}^n |x_i - y_i|$, where $\nu = \max_i \|\mathbf{e}_i\|$. For each $\epsilon > 0$, set $\delta = \epsilon/n\nu$. If $|x_i - y_i| < \delta$ for each i, then, using (5.1.6), $\big|\ \|\mathbf{x}\| - \|\mathbf{y}\|\ \big| \leq \|\mathbf{x} - \mathbf{y}\| < \nu n \delta = \epsilon$.

5.1.8. To show that $\|\mathbf{x}\|_1 \leq \sqrt{n}\,\|\mathbf{x}\|_2$, apply the CBS inequality to the standard inner product of a vector of all 1's with a vector whose components are the $|x_i|$'s.

5.1.9. If $\mathbf{y} = \alpha\mathbf{x}$, then $|\mathbf{x}^*\mathbf{y}| = |\alpha|\ \|\mathbf{x}\|^2 = \|\mathbf{x}\|\ \|\mathbf{y}\|$. Conversely, if $|\mathbf{x}^*\mathbf{y}| = \|\mathbf{x}\|\ \|\mathbf{y}\|$, then (5.1.4) implies that $\|\alpha\mathbf{x} - \mathbf{y}\| = 0$, and hence $\alpha\mathbf{x} - \mathbf{y} = \mathbf{0}$—recall (5.1.1).

5.1.10. If $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha > 0$, then $\|\mathbf{x} + \mathbf{y}\| = \|(1 + \alpha)\mathbf{x}\| = (1 + \alpha)\,\|\mathbf{x}\| = \|\mathbf{x}\| + \|\mathbf{y}\|$. Conversely, $\|\mathbf{x} + \mathbf{y}\| = \|\mathbf{x}\| + \|\mathbf{y}\| \implies (\|\mathbf{x}\| + \|\mathbf{y}\|)^2 = \|\mathbf{x} + \mathbf{y}\|^2 \implies$

$$\begin{aligned} \|\mathbf{x}\|^2 + 2\,\|\mathbf{x}\|\ \|\mathbf{y}\| + \|\mathbf{y}\|^2 &= (\mathbf{x}^* + \mathbf{y}^*)\,(\mathbf{x} + \mathbf{y}) \\ &= \mathbf{x}^*\mathbf{x} + \mathbf{x}^*\mathbf{y} + \mathbf{y}^*\mathbf{x} + \mathbf{y}^*\mathbf{y} \\ &= \|\mathbf{x}\|^2 + 2\,\mathrm{Re}(\mathbf{x}^*\mathbf{y}) + \|\mathbf{y}\|^2, \end{aligned}$$

and hence $\|\mathbf{x}\|\ \|\mathbf{y}\| = \mathrm{Re}\,(\mathbf{x}^*\mathbf{y})$. But it's always true that $\mathrm{Re}\,(\mathbf{x}^*\mathbf{y}) \leq \left|\mathbf{x}^*\mathbf{y}\right|$, so the CBS inequality yields

$$\|\mathbf{x}\|\ \|\mathbf{y}\| = \mathrm{Re}\,(\mathbf{x}^*\mathbf{y}) \leq \left|\mathbf{x}^*\mathbf{y}\right| \leq \|\mathbf{x}\|\ \|\mathbf{y}\|.$$

In other words, $\left|\mathbf{x}^*\mathbf{y}\right| = \|\mathbf{x}\|\ \|\mathbf{y}\|$. We know from Exercise 5.1.9 that equality in the CBS inequality implies $\mathbf{y} = \alpha\mathbf{x}$, where $\alpha = \mathbf{x}^*\mathbf{y}/\mathbf{x}^*\mathbf{x}$. We now need to show that this α is real and positive. Using $\mathbf{y} = \alpha\mathbf{x}$ in the equality $\|\mathbf{x} + \mathbf{y}\| =$

$\|\mathbf{x}\| + \|\mathbf{y}\|$ produces $|1+\alpha| = 1+|\alpha|$, or $|1+\alpha|^2 = (1+|\alpha|)^2$. Expanding this yields

$$\begin{aligned}(1+\bar{\alpha})(1+\alpha) &= 1 + 2|\alpha| + |\alpha|^2 \\ \Longrightarrow \quad 1 + 2\,\mathrm{Re}(\alpha) + \bar{\alpha}\alpha &= 1 + 2|\alpha| + \bar{\alpha}\alpha \\ \Longrightarrow \quad \mathrm{Re}(\alpha) &= |\alpha|,\end{aligned}$$

which implies that α must be real. Furthermore, $\alpha = \mathrm{Re}\,(\alpha) = |\alpha| \geq 0$. Since $\mathbf{y} = \alpha\mathbf{x}$ and $\mathbf{y} \neq \mathbf{0}$, it follows that $\alpha \neq 0$, and therefore $\alpha > 0$.

5.1.11. This is a consequence of Hölder's inequality because

$$|\mathbf{x}^T\mathbf{y}| = |\mathbf{x}^T(\mathbf{y} - \alpha\mathbf{e})| \leq \|\mathbf{x}\|_1 \, \|\mathbf{y} - \alpha\mathbf{e}\|_\infty$$

for all α, and $\min_\alpha \|\mathbf{y} - \alpha\mathbf{e}\|_\infty = (y_{\max} - y_{\min})/2$ (with the minimum being attained at $\alpha = (y_{\max} + y_{\min})/2$).

5.1.12. (a) It's not difficult to see that $f'(t) < 0$ for $t < 1$, and $f'(t) > 0$ for $t > 1$, so we can conclude that $f(t) > f(1) = 0$ for $t \neq 1$. The desired inequality follows by setting $t = \alpha/\beta$.

(b) This inequality follows from the inequality of part (a) by setting

$$\alpha = |\hat{x}_i|^p, \quad \beta = |\hat{y}_i|^q, \quad \lambda = 1/p, \quad \text{and} \quad (1-\lambda) = 1/q.$$

(c) Hölder's inequality results from part (b) by setting $\hat{x}_i = x_i / \|\mathbf{x}\|_p$ and $\hat{y}_i = y_i / \|\mathbf{y}\|_q$. To obtain the "vector form" of the inequality, use the triangle inequality for complex numbers to write

$$\begin{aligned}|\mathbf{x}^*\mathbf{y}| &= \left|\sum_{i=1}^n \overline{x_i} y_i\right| \leq \sum_{i=1}^n |\overline{x_i}|\,|y_i| = \sum_{i=1}^n |x_i y_i| \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q} \\ &= \|\mathbf{x}\|_p \, \|\mathbf{y}\|_q .\end{aligned}$$

5.1.13. For $p = 1$, Minkowski's inequality is a consequence of the triangle inequality for scalars. The inequality in the hint follows from the fact that $p = 1 + p/q$ together with the scalar triangle inequality, and it implies that

$$\sum_{i=1}^n |x_i + y_i|^p = \sum_{i=1}^n |x_i + y_i|\,|x_i + y_i|^{p/q} \leq \sum_{i=1}^n |x_i|\,|x_i + y_i|^{p/q} + \sum_{i=1}^n |y_i|\,|x_i + y_i|^{p/q}.$$

Application of Hölder's inequality produces

$$\begin{aligned}\sum_{i=1}^n |x_i|\,|x_i + y_i|^{p/q} &\leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |x_i + y_i|^p\right)^{1/q} \\ &= \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |x_i + y_i|^p\right)^{(p-1)/p} \\ &= \|\mathbf{x}\|_p \, \|\mathbf{x} + \mathbf{y}\|_p^{p-1} .\end{aligned}$$

Similarly, $\displaystyle\sum_{i=1}^{n} |y_i|\,|x_i+y_i|^{p/q} \le \|\mathbf{y}\|_p\,\|\mathbf{x}+\mathbf{y}\|_p^{p-1}$, and therefore

$$\|\mathbf{x}+\mathbf{y}\|_p^p \le \Big(\|\mathbf{x}\|_p + \|\mathbf{y}\|_p\Big)\,\|\mathbf{x}+\mathbf{y}\|_p^{p-1} \quad\Longrightarrow\quad \|\mathbf{x}+\mathbf{y}\|_p \le \|\mathbf{x}\|_p + \|\mathbf{y}\|_p\,.$$

Solutions for exercises in section 5. 2

5.2.1. $\|\mathbf{A}\|_F = \left[\sum_{i,j}|a_{ij}|^2\right]^{1/2} = [trace\,(\mathbf{A}^*\mathbf{A})]^{1/2} = \sqrt{10}$,

$\|\mathbf{B}\|_F = \sqrt{3}$, and $\|\mathbf{C}\|_F = \sqrt{9}$.

5.2.2. (a) $\|\mathbf{A}\|_1 =$ max absolute column sum $= 4$, and $\|\mathbf{A}\|_\infty =$ max absolute row sum $= 3$. $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}}$, where $\lambda_{\max}$ is the largest value of λ for which $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ is singular. Determine these λ's by row reduction.

$$\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I} = \begin{pmatrix} 2 - -\lambda & -4 \\ -4 & 8-\lambda \end{pmatrix} \longrightarrow \begin{pmatrix} -4 & 8-\lambda \\ 2-\lambda & -4 \end{pmatrix}$$
$$\longrightarrow \begin{pmatrix} -4 & 8-\lambda \\ 0 & -4 + \frac{2-\lambda}{4}(8-\lambda) \end{pmatrix}$$

This matrix is singular if and only if the second pivot is zero, so we must have $(2-\lambda)(8-\lambda) - 16 = 0 \implies \lambda^2 - 10\lambda = 0 \implies \lambda = 0,\ \lambda = 10$, and therefore $\|\mathbf{A}\|_2 = \sqrt{10}$.

(b) Use the same technique to get $\|\mathbf{B}\|_1 = \|\mathbf{B}\|_2 = \|\mathbf{B}\|_\infty = 1$, and

(c) $\|\mathbf{C}\|_1 = \|\mathbf{C}\|_\infty = 10$ and $\|\mathbf{C}\|_2 = 9$.

5.2.3. (a) $\|\mathbf{I}\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{I}\mathbf{x}\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{x}\| = 1$.

(b) $\|\mathbf{I}_{n\times n}\|_F = \left[trace\,\left(\mathbf{I}^T\mathbf{I}\right)\right]^{1/2} = \sqrt{n}$.

5.2.4. Use the fact that $trace\,(\mathbf{AB}) = trace\,(\mathbf{BA})$ (recall Example 3.6.5) to write

$$\|\mathbf{A}\|_F^2 = trace\,(\mathbf{A}^*\mathbf{A}) = trace\,(\mathbf{A}\mathbf{A}^*) = \|\mathbf{A}^*\|_F^2\,.$$

5.2.5. (a) For $\mathbf{x} = \mathbf{0}$, the statement is trivial. For $\mathbf{x} \ne \mathbf{0}$, we have $\|(\mathbf{x}/\|\mathbf{x}\|)\| = 1$, so for any particular $\mathbf{x}_0 \ne \mathbf{0}$,

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{x}\| = \max_{\mathbf{x}\ne\mathbf{0}}\left\|\mathbf{A}\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| \ge \frac{\|\mathbf{A}\mathbf{x}_0\|}{\|\mathbf{x}_0\|} \quad\Longrightarrow\quad \|\mathbf{A}\mathbf{x}_0\| \le \|\mathbf{A}\|\,\|\mathbf{x}_0\|\,.$$

(b) Let $\mathbf{x}_0$ be a vector such that $\|\mathbf{x}_0\| = 1$ and

$$\|\mathbf{A}\mathbf{B}\mathbf{x}_0\| = \max_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{B}\mathbf{x}\| = \|\mathbf{A}\mathbf{B}\|\,.$$

Make use of the result of part (a) to write

$$\|\mathbf{A}\mathbf{B}\| = \|\mathbf{A}\mathbf{B}\mathbf{x}_0\| \le \|\mathbf{A}\|\,\|\mathbf{B}\mathbf{x}_0\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|\,\|\mathbf{x}_0\| = \|\mathbf{A}\|\,\|\mathbf{B}\|\,.$$

(c) $\|\mathbf{A}\| = \max\limits_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| \leq \max\limits_{\|\mathbf{x}\|\leq 1} \|\mathbf{A}\mathbf{x}\|$ because $\{\mathbf{x} \mid \|\mathbf{x}\| = 1\} \subset \{\mathbf{x} \mid \|\mathbf{x}\| \leq 1\}$. If there would exist a vector $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| < 1$ and $\|\mathbf{A}\| < \|\mathbf{A}\mathbf{x}_0\|$, then part (a) would insure that $\|\mathbf{A}\| < \|\mathbf{A}\mathbf{x}_0\| \leq \|\mathbf{A}\| \, \|\mathbf{x}_0\| < \|\mathbf{A}\|$, which is impossible.

5.2.6. (a) Applying the CBS inequality yields

$$|\mathbf{y}^*\mathbf{A}\mathbf{x}| \leq \|\mathbf{y}\|_2 \, \|\mathbf{A}\mathbf{x}\|_2 \implies \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{y}^*\mathbf{A}\mathbf{x}| \leq \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2 .$$

Now show that equality is actually attained for some pair $\mathbf{x}$ and $\mathbf{y}$ on the unit 2-sphere. To do so, notice that if $\mathbf{x}_0$ is a vector of unit length such that

$$\|\mathbf{A}\mathbf{x}_0\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2 , \quad \text{and if} \quad \mathbf{y}_0 = \frac{\mathbf{A}\mathbf{x}_0}{\|\mathbf{A}\mathbf{x}_0\|_2} = \frac{\mathbf{A}\mathbf{x}_0}{\|\mathbf{A}\|_2},$$

then

$$\mathbf{y}_0^*\mathbf{A}\mathbf{x}_0 = \frac{\mathbf{x}_0^*\mathbf{A}^*\mathbf{A}\mathbf{x}_0}{\|\mathbf{A}\|_2} = \frac{\|\mathbf{A}\mathbf{x}_0\|_2^2}{\|\mathbf{A}\|_2} = \frac{\|\mathbf{A}\|_2^2}{\|\mathbf{A}\|_2} = \|\mathbf{A}\|_2 .$$

(b) This follows directly from the result of part (a) because

$$\|\mathbf{A}\|_2 = \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{y}^*\mathbf{A}\mathbf{x}| = \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |(\mathbf{y}^*\mathbf{A}\mathbf{x})^*| = \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{x}^*\mathbf{A}^*\mathbf{y}| = \|\mathbf{A}^*\|_2 .$$

(c) Use part (a) with the CBS inequality to write

$$\|\mathbf{A}^*\mathbf{A}\|_2 = \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} |\mathbf{y}^*\mathbf{A}^*\mathbf{A}\mathbf{x}| \leq \max_{\substack{\|\mathbf{x}\|_2=1 \\ \|\mathbf{y}\|_2=1}} \|\mathbf{A}\mathbf{y}\|_2 \, \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2^2 .$$

To see that equality is attained, let $\mathbf{x} = \mathbf{y} = \mathbf{x}_0$, where $\mathbf{x}_0$ is a vector of unit length such that $\|\mathbf{A}\mathbf{x}_0\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2$, and observe

$$|\mathbf{x}_0^*\mathbf{A}^*\mathbf{A}\mathbf{x}_0| = \mathbf{x}_0^*\mathbf{A}^*\mathbf{A}\mathbf{x}_0 = \|\mathbf{A}\mathbf{x}_0\|_2^2 = \|\mathbf{A}\|_2^2 .$$

(d) Let $\mathbf{D} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$. We know from (5.2.7) that $\|\mathbf{D}\|_2^2$ is the largest value λ such that $\mathbf{D}^T\mathbf{D} - \lambda\mathbf{I}$ is singular. But $\mathbf{D}^T\mathbf{D} - \lambda\mathbf{I}$ is singular if and only if $\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}$ or $\mathbf{B}^T\mathbf{B} - \lambda\mathbf{I}$ is singular, so $\lambda_{\max}(\mathbf{D}) = \max\{\lambda_{\max}(\mathbf{A}), \lambda_{\max}(\mathbf{B})\}$.

(e) If $\mathbf{U}\mathbf{U}^* = \mathbf{I}$, then $\|\mathbf{U}^*\mathbf{A}\mathbf{x}\|_2^2 = \mathbf{x}^*\mathbf{A}^*\mathbf{U}\mathbf{U}^*\mathbf{A}\mathbf{x} = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2$, so $\|\mathbf{U}^*\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{U}^*\mathbf{A}\mathbf{x}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\|_2$. Now, if $\mathbf{V}^*\mathbf{V} = \mathbf{I}$, use what was just established with part (b) to write

$$\|\mathbf{A}\mathbf{V}\|_2 = \|(\mathbf{A}\mathbf{V})^*\|_2 = \|\mathbf{V}^*\mathbf{A}^*\|_2 = \|\mathbf{A}^*\|_2 = \|\mathbf{A}\|_2 \implies \|\mathbf{U}^*\mathbf{A}\mathbf{V}\|_2 = \|\mathbf{A}\|_2.$$

5.2.7. Proceed as follows.

$$\frac{1}{\min\limits_{\|\mathbf{x}\|=1}\left\|\mathbf{A}^{-1}\mathbf{x}\right\|} = \max_{\|\mathbf{x}\|=1}\left\{\frac{1}{\|\mathbf{A}^{-1}\mathbf{x}\|}\right\} = \max_{\mathbf{y}\neq\mathbf{0}}\left\{\frac{1}{\left\|\mathbf{A}^{-1}\frac{(\mathbf{A}\mathbf{y})}{\|\mathbf{A}\mathbf{y}\|}\right\|}\right\}$$

$$= \max_{\mathbf{y}\neq\mathbf{0}}\frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{A}^{-1}(\mathbf{A}\mathbf{y})\|} = \max_{\mathbf{y}\neq\mathbf{0}}\frac{\|\mathbf{A}\mathbf{y}\|}{\|\mathbf{y}\|} = \max_{\mathbf{y}\neq\mathbf{0}}\left\|\mathbf{A}\left(\frac{\mathbf{y}}{\|\mathbf{y}\|}\right)\right\|$$

$$= \max_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{x}\| = \|\mathbf{A}\|$$

5.2.8. Use (5.2.6) on p. 280 to write $\|(z\mathbf{I}-\mathbf{A})^{-1}\| = (1/\min_{\|\mathbf{x}\|=1}\|(z\mathbf{I}-\mathbf{A})\mathbf{x}\|)$, and let $\mathbf{w}$ be a vector for which $\|\mathbf{w}\| = 1$ and $\|(z\mathbf{I}-\mathbf{A})\mathbf{w}\| = \min_{\|\mathbf{x}\|=1}\|(z\mathbf{I}-\mathbf{A})\mathbf{x}\|$. Use $\|\mathbf{A}\mathbf{w}\| \leq \|\mathbf{A}\| < |z|$ together with the "backward triangle inequality" from Example 5.1.1 (p. 273) to write

$$\|(z\mathbf{I}-\mathbf{A})\mathbf{w}\| = \|z\mathbf{w}-\mathbf{A}\mathbf{w}\| \geq \big|\|z\mathbf{w}\| - \|\mathbf{A}\mathbf{w}\|\big| = \big||z| - \|\mathbf{A}\mathbf{w}\|\big|$$
$$= |z| - \|\mathbf{A}\mathbf{w}\| \geq |z| - \|\mathbf{A}\|.$$

Consequently, $\min_{\|\mathbf{x}\|=1}\|(z\mathbf{I}-\mathbf{A})\mathbf{x}\| = \|(z\mathbf{I}-\mathbf{A})\mathbf{w}\| \geq |z| - \|\mathbf{A}\|$ implies that

$$\|(z\mathbf{I}-\mathbf{A})^{-1}\| = \frac{1}{\min\limits_{\|\mathbf{x}\|=1}\|(z\mathbf{I}-\mathbf{A})\mathbf{x}\|} \leq \frac{1}{|z| - \|\mathbf{A}\|}.$$

Solutions for exercises in section 5. 3

5.3.1. Only (c) is an inner product. The expressions in (a) and (b) each fail the first condition of the definition (5.3.1), and (d) fails the second.

5.3.2. (a) $\langle\mathbf{x}|\mathbf{y}\rangle = 0 \;\; \forall\; \mathbf{x}\in\mathcal{V} \implies \langle\mathbf{y}|\mathbf{y}\rangle = 0 \implies \mathbf{y} = \mathbf{0}$.

(b) $\langle\alpha\mathbf{x}|\mathbf{y}\rangle = \overline{\langle\mathbf{y}|\alpha\mathbf{x}\rangle} = \overline{\alpha\,\langle\mathbf{y}|\mathbf{x}\rangle} = \overline{\alpha}\overline{\langle\mathbf{y}|\mathbf{x}\rangle} = \overline{\alpha}\,\langle\mathbf{x}|\mathbf{y}\rangle$

(c) $\langle\mathbf{x}+\mathbf{y}|\mathbf{z}\rangle = \overline{\langle\mathbf{z}|\mathbf{x}+\mathbf{y}\rangle} = \overline{\langle\mathbf{z}|\mathbf{x}\rangle + \langle\mathbf{z}|\mathbf{y}\rangle} = \overline{\langle\mathbf{z}|\mathbf{x}\rangle} + \overline{\langle\mathbf{z}|\mathbf{y}\rangle} = \langle\mathbf{x}|\mathbf{z}\rangle + \langle\mathbf{y}|\mathbf{z}\rangle$

5.3.3. The first property in (5.2.3) holds because $\langle\mathbf{x}|\mathbf{x}\rangle \geq 0$ for all $\mathbf{x}\in\mathcal{V}$ implies $\|\mathbf{x}\| = \sqrt{\langle\mathbf{x}|\mathbf{x}\rangle} \geq 0$, and $\|\mathbf{x}\| = 0 \iff \langle\mathbf{x}|\mathbf{x}\rangle = 0 \iff \mathbf{x} = \mathbf{0}$. The second property in (5.2.3) holds because

$$\|\alpha\mathbf{x}\|^2 = \langle\alpha\mathbf{x}|\alpha\mathbf{x}\rangle = \alpha\,\langle\alpha\mathbf{x}|\mathbf{x}\rangle = \alpha\overline{\langle\mathbf{x}|\alpha\mathbf{x}\rangle} = \alpha\overline{\alpha}\overline{\langle\mathbf{x}|\mathbf{x}\rangle} = |\alpha|^2\,\langle\mathbf{x}|\mathbf{x}\rangle = |\alpha|^2\,\|\mathbf{x}\|^2.$$

5.3.4. $0 \leq \|\mathbf{x}-\mathbf{y}\|^2 = \langle\mathbf{x}-\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle = \langle\mathbf{x}|\mathbf{x}\rangle - 2\,\langle\mathbf{x}|\mathbf{y}\rangle + \langle\mathbf{y}|\mathbf{y}\rangle = \|\mathbf{x}\|^2 - 2\,\langle\mathbf{x}|\mathbf{y}\rangle + \|\mathbf{y}\|^2$

5.3.5. (a) Use the CBS inequality with the Frobenius matrix norm and the standard inner product as illustrated in Example 5.3.3, and set $\mathbf{A} = \mathbf{I}$.

(b) Proceed as in part (a), but this time set $\mathbf{A} = \mathbf{B}^T$ (recall from Example 3.6.5 that $trace\left(\mathbf{B}^T\mathbf{B}\right) = trace\left(\mathbf{B}\mathbf{B}^T\right)$).

(c) Use the result of Exercise 5.3.4 with the Frobenius matrix norm and the inner product for matrices.

5.3.6. Suppose that parallelogram identity holds, and verify that (5.3.10) satisfies the four conditions in (5.3.1). The first condition follows because $\langle \mathbf{x}|\mathbf{x}\rangle_r = \|\mathbf{x}\|^2$ and $\langle \mathbf{ix}|\mathbf{x}\rangle_r = 0$ combine to yield $\langle \mathbf{x}|\mathbf{x}\rangle = \|\mathbf{x}\|^2$. The second condition (for real α) and third condition hold by virtue of the argument for (5.3.7). We will prove the fourth condition and then return to show that the second holds for complex α. By observing that $\langle \mathbf{x}|\mathbf{y}\rangle_r = \langle \mathbf{y}|\mathbf{x}\rangle_r$ and $\langle \mathbf{ix}|\mathbf{iy}\rangle_r = \langle \mathbf{x}|\mathbf{y}\rangle_r$, we have

$$\langle \mathbf{iy}|\mathbf{x}\rangle_r = \big\langle \mathbf{iy}\big|-\mathrm{i}^2\mathbf{x}\big\rangle_r = \langle \mathbf{y}|-\mathbf{ix}\rangle_r = -\langle \mathbf{y}|\mathbf{ix}\rangle_r = -\langle \mathbf{ix}|\mathbf{y}\rangle_r,$$

and hence

$$\langle \mathbf{y}|\mathbf{x}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle_r + \mathrm{i}\langle \mathbf{iy}|\mathbf{x}\rangle_r = \langle \mathbf{y}|\mathbf{x}\rangle_r - \mathrm{i}\langle \mathbf{ix}|\mathbf{y}\rangle_r = \langle \mathbf{x}|\mathbf{y}\rangle_r - \mathrm{i}\langle \mathbf{ix}|\mathbf{y}\rangle_r = \overline{\langle \mathbf{x}|\mathbf{y}\rangle}.$$

Now prove that $\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle$ for all complex α. Begin by showing it is true for $\alpha = \mathrm{i}$.

$$\begin{aligned}\langle \mathbf{x}|\mathbf{iy}\rangle &= \langle \mathbf{x}|\mathbf{iy}\rangle_r + \mathrm{i}\langle \mathbf{ix}|\mathbf{iy}\rangle_r = \langle \mathbf{x}|\mathbf{iy}\rangle_r + \mathrm{i}\langle \mathbf{x}|\mathbf{y}\rangle_r = \langle \mathbf{iy}|\mathbf{x}\rangle_r + \mathrm{i}\langle \mathbf{x}|\mathbf{y}\rangle_r \\ &= -\langle \mathbf{ix}|\mathbf{y}\rangle_r + \mathrm{i}\langle \mathbf{x}|\mathbf{y}\rangle_r = \mathrm{i}\left(\langle \mathbf{x}|\mathbf{y}\rangle_r + \mathrm{i}\langle \mathbf{ix}|\mathbf{y}\rangle_r\right) \\ &= \mathrm{i}\langle \mathbf{x}|\mathbf{y}\rangle\end{aligned}$$

For $\alpha = \xi + \mathrm{i}\eta$,

$$\langle \mathbf{x}|\alpha\mathbf{y}\rangle = \langle \mathbf{x}|\xi\mathbf{y} + \mathrm{i}\eta\mathbf{y}\rangle = \langle \mathbf{x}|\xi\mathbf{y}\rangle + \langle \mathbf{x}|\mathrm{i}\eta\mathbf{y}\rangle = \xi\langle \mathbf{x}|\mathbf{y}\rangle + \mathrm{i}\eta\langle \mathbf{x}|\mathbf{y}\rangle = \alpha\langle \mathbf{x}|\mathbf{y}\rangle.$$

Conversely, if $\langle \star|\star\rangle$ is any inner product on $\mathcal{V}$, then with $\|\star\|^2 = \langle \star|\star\rangle$ we have

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle + \langle \mathbf{x}-\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle \\ &= \|\mathbf{x}\|^2 + 2\mathrm{Re}\langle \mathbf{x}|\mathbf{y}\rangle + \|\mathbf{y}\|^2 + \|\mathbf{x}\|^2 - 2\mathrm{Re}\langle \mathbf{x}|\mathbf{y}\rangle + \|\mathbf{y}\|^2 \\ &= 2\left(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\right).\end{aligned}$$

5.3.7. The parallelogram identity (5.3.7) fails to hold for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}^n$. For example, if $\mathbf{x} = \mathbf{e}_1$ and $\mathbf{y} = \mathbf{e}_2$, then

$$\|\mathbf{e}_1 + \mathbf{e}_2\|_\infty^2 + \|\mathbf{e}_1 - \mathbf{e}_2\|_\infty^2 = 2, \quad \text{but} \quad 2\big(\|\mathbf{e}_1\|_\infty^2 + \|\mathbf{e}_2\|_\infty^2\big) = 4.$$

5.3.8. (a) As shown in Example 5.3.2, the Frobenius matrix norm $\mathcal{C}^{n\times n}$ is generated by the standard matrix inner product (5.3.2), so the result on p. 290 guarantees that $\|\star\|_F$ satisfies the parallelogram identity.

5.3.9. No, because the parallelogram inequality (5.3.7) doesn't hold. To see that $\|\mathbf{X}+\mathbf{Y}\|^2 + \|\mathbf{X}-\mathbf{Y}\|^2 = 2\big(\|\mathbf{X}\|^2 + \|\mathbf{Y}\|^2\big)$ is not valid for all $\mathbf{X}, \mathbf{Y} \in \mathcal{C}^{n\times n}$, let $\mathbf{X} = \mathrm{diag}\,(1, 0, \ldots, 0)$ and $\mathbf{Y} = \mathrm{diag}\,(0, 1, \ldots, 0)$. For $\star = 1, 2,$ or ∞,

$$\|\mathbf{X}+\mathbf{Y}\|_\star^2 + \|\mathbf{X}-\mathbf{Y}\|_\star^2 = 1 + 1 = 2, \quad \text{but} \quad 2\big(\|\mathbf{X}\|_\star^2 + \|\mathbf{Y}\|_\star^2\big) = 4.$$

Solutions for exercises in section 5. 4

5.4.1. (a), (b), and (e) are orthogonal pairs.

5.4.2. First find $\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ such that $3\alpha_1 - 2\alpha_2 = 0$, and then normalize $\mathbf{v}$. The second must be the negative of $\mathbf{v}$.

5.4.3. (a) Simply verify that $\mathbf{x}_i^T\mathbf{x}_j = 0$ for $i \neq j$.

(b) Let $\mathbf{x}_4^T = (\,\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \alpha_4\,)$, and notice that $\mathbf{x}_i^T\mathbf{x}_4 = 0$ for $i = 1, 2, 3$ is three homogeneous equations in four unknowns

$$\begin{pmatrix} 1 & -1 & 0 & 2 \\ 1 & 1 & 1 & 0 \\ -1 & -1 & 2 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \implies \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \beta \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

(c) Simply normalize the set by dividing each vector by its norm.

5.4.4. The Fourier coefficients are

$$\xi_1 = \langle \mathbf{u}_1 | \mathbf{x} \rangle = \frac{1}{\sqrt{2}}, \quad \xi_2 = \langle \mathbf{u}_2 | \mathbf{x} \rangle = \frac{-1}{\sqrt{3}}, \quad \xi_3 = \langle \mathbf{u}_3 | \mathbf{x} \rangle = \frac{-5}{\sqrt{6}},$$

so

$$\mathbf{x} = \xi_1\mathbf{u}_1 + \xi_2\mathbf{u}_2 + \xi_3\mathbf{u}_3 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{5}{6}\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

5.4.5. If $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{U}_3$, and $\mathbf{U}_4$ denote the elements of $\mathcal{B}$, verify they constitute an orthonormal set by showing that

$$\langle \mathbf{U}_i | \mathbf{U}_j \rangle = trace(\mathbf{U}_i^T\mathbf{U}_j) = 0 \text{ for } i \neq j \quad \text{and} \quad \|\mathbf{U}_i\| = \sqrt{trace(\mathbf{U}_i^T\mathbf{U}_i)} = 1.$$

Consequently, $\mathcal{B}$ is linearly independent—recall (5.4.2)—and therefore $\mathcal{B}$ is a basis because it is a *maximal* independent set—part (b) of Exercise 4.4.4 insures $\dim \Re^{2\times 2} = 4$. The Fourier coefficients $\langle \mathbf{U}_i | \mathbf{A} \rangle = trace(\mathbf{U}_i^T\mathbf{A})$ are

$$\langle \mathbf{U}_1 | \mathbf{A} \rangle = \frac{2}{\sqrt{2}}, \quad \langle \mathbf{U}_2 | \mathbf{A} \rangle = 0, \quad \langle \mathbf{U}_3 | \mathbf{A} \rangle = 1, \quad \langle \mathbf{U}_4 | \mathbf{A} \rangle = 1,$$

so the Fourier expansion of $\mathbf{A}$ is $\mathbf{A} = (2/\sqrt{2})\mathbf{U}_1 + \mathbf{U}_3 + \mathbf{U}_4$.

5.4.6. $\cos\theta = \mathbf{x}^T\mathbf{y} / \|\mathbf{x}\|\,\|\mathbf{y}\| = 1/2$, so $\theta = \pi/3$.

5.4.7. This follows because each vector has a unique representation in terms of a basis—see Exercise 4.4.8 or the discussion of coordinates in §4.7.

5.4.8. If the columns of $\mathbf{U} = [\mathbf{u}_1 \,|\, \mathbf{u}_2 \,|\, \cdots \,|\, \mathbf{u}_n]$ are an orthonormal basis for $\mathcal{C}^n$, then

$$[\mathbf{U}^*\mathbf{U}]_{ij} = \mathbf{u}_i^*\mathbf{u}_j = \begin{cases} 1 & \text{when } i = j, \\ 0 & \text{when } i \neq j, \end{cases} \qquad (\ddagger)$$

and, therefore, $\mathbf{U}^*\mathbf{U} = \mathbf{I}$. Conversely, if $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, then (‡) holds, so the columns of $\mathbf{U}$ are orthonormal—they are a basis for $\mathcal{C}^n$ because orthonormal sets are always linearly independent.

5.4.9. Equations (4.5.5) and (4.5.6) guarantee that

$$R(\mathbf{A}) = R(\mathbf{A}\mathbf{A}^*) \quad \text{and} \quad N(\mathbf{A}) = N(\mathbf{A}^*\mathbf{A}),$$

and consequently $\mathbf{r} \in R(\mathbf{A}) = R(\mathbf{A}\mathbf{A}^*) \implies \mathbf{r} = \mathbf{A}\mathbf{A}^*\mathbf{x}$ for some $\mathbf{x}$, and $\mathbf{n} \in N(\mathbf{A}) = N(\mathbf{A}^*\mathbf{A}) \implies \mathbf{A}^*\mathbf{A}\mathbf{n} = \mathbf{0}$. Therefore,

$$\langle \mathbf{r}|\mathbf{n}\rangle = \mathbf{r}^*\mathbf{n} = \mathbf{x}^*\mathbf{A}\mathbf{A}^*\mathbf{n} = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{n} = \mathbf{0}.$$

5.4.10. (a) $\pi/4$ (b) $\pi/2$

5.4.11. The number $\mathbf{x}^T\mathbf{y}$ or $\mathbf{x}^*\mathbf{y}$ will in general be complex. In order to guarantee that we end up with a real number, we should take

$$\cos\theta = \frac{|\mathrm{Re}\,(\mathbf{x}^*\mathbf{y})\,|}{\|\mathbf{x}\|\ \|\mathbf{y}\|}.$$

5.4.12. Use the Fourier expansion $\mathbf{y} = \sum_i \langle \mathbf{u}_i|\mathbf{y}\rangle\,\mathbf{u}_i$ together with the various properties of an inner product to write

$$\langle \mathbf{x}|\mathbf{y}\rangle = \left\langle \mathbf{x} \middle| \sum_i \langle \mathbf{u}_i|\mathbf{y}\rangle\,\mathbf{u}_i \right\rangle = \sum_i \langle \mathbf{x}|\,\langle \mathbf{u}_i|\mathbf{y}\rangle\,\mathbf{u}_i\rangle = \sum_i \langle \mathbf{u}_i|\mathbf{y}\rangle\,\langle \mathbf{x}|\mathbf{u}_i\rangle\,.$$

5.4.13. In a real space, $\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle$, so the third condition in the definition (5.3.1) of an inner product and Exercise 5.3.2(c) produce

$$\begin{aligned}\langle \mathbf{x}+\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle &= \langle \mathbf{x}+\mathbf{y}|\mathbf{x}\rangle - \langle \mathbf{x}+\mathbf{y}|\mathbf{y}\rangle \\ &= \langle \mathbf{x}|\mathbf{x}\rangle + \langle \mathbf{y}|\mathbf{x}\rangle - \langle \mathbf{x}|\mathbf{y}\rangle - \langle \mathbf{y}|\mathbf{y}\rangle \\ &= \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 = 0.\end{aligned}$$

5.4.14. (a) In a real space, $\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle$, so the third condition in the definition (5.3.1) of an inner product and Exercise 5.3.2(c) produce

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle = \langle \mathbf{x}+\mathbf{y}|\mathbf{x}\rangle + \langle \mathbf{x}+\mathbf{y}|\mathbf{y}\rangle \\ &= \langle \mathbf{x}|\mathbf{x}\rangle + \langle \mathbf{y}|\mathbf{x}\rangle + \langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{y}|\mathbf{y}\rangle \\ &= \|\mathbf{x}\|^2 + 2\,\langle \mathbf{x}|\mathbf{y}\rangle + \|\mathbf{y}\|^2\,,\end{aligned}$$

and hence $\langle \mathbf{x}|\mathbf{y}\rangle = 0$ if and only if $\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.

(b) In a complex space, $\mathbf{x} \perp \mathbf{y} \implies \|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, but the converse is not valid—e.g., consider $\mathcal{C}^2$ with the standard inner product, and let $\mathbf{x} = \begin{pmatrix} -\mathrm{i} \\ 1 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 1 \\ \mathrm{i} \end{pmatrix}$.

(c) Again, using the properties of a general inner product, derive the expansion

$$\begin{aligned}\|\alpha\mathbf{x}+\beta\mathbf{y}\|^2 &= \langle\alpha\mathbf{x}+\beta\mathbf{y}|\alpha\mathbf{x}+\beta\mathbf{y}\rangle\\ &= \langle\alpha\mathbf{x}|\alpha\mathbf{x}\rangle+\langle\alpha\mathbf{x}|\beta\mathbf{y}\rangle+\langle\beta\mathbf{y}|\alpha\mathbf{x}\rangle+\langle\beta\mathbf{y}|\beta\mathbf{y}\rangle\\ &= \|\alpha\mathbf{x}\|^2+\overline{\alpha}\beta\langle\mathbf{x}|\mathbf{y}\rangle+\overline{\beta}\alpha\langle\mathbf{y}|\mathbf{x}\rangle+\|\beta\mathbf{y}\|^2.\end{aligned}$$

Clearly, $\mathbf{x}\perp\mathbf{y} \implies \|\alpha\mathbf{x}+\beta\mathbf{y}\|^2 = \|\alpha\mathbf{x}\|^2+\|\beta\mathbf{y}\|^2 \;\forall\; \alpha,\beta$. Conversely, if $\|\alpha\mathbf{x}+\beta\mathbf{y}\|^2 = \|\alpha\mathbf{x}\|^2+\|\beta\mathbf{y}\|^2 \;\forall\; \alpha,\beta$, then $\overline{\alpha}\beta\langle\mathbf{x}|\mathbf{y}\rangle+\overline{\beta}\alpha\langle\mathbf{y}|\mathbf{x}\rangle = 0 \;\forall\; \alpha,\beta$. Letting $\alpha=\langle\mathbf{x}|\mathbf{y}\rangle$ and $\beta=1$ produces the conclusion that $2|\langle\mathbf{x}|\mathbf{y}\rangle|^2=0$, and thus $\langle\mathbf{x}|\mathbf{y}\rangle=0$.

5.4.15. (a) $\cos\theta_i = \langle\mathbf{u}_i|\mathbf{x}\rangle/\|\mathbf{u}_i\|\,\|\mathbf{x}\| = \langle\mathbf{u}_i|\mathbf{x}\rangle/\|\mathbf{x}\| = \xi_i/\|\mathbf{x}\|$

(b) Use the Pythagorean theorem (Exercise 5.4.14) to write

$$\begin{aligned}\|\mathbf{x}\|^2 &= \|\xi_1\mathbf{u}_1+\xi_2\mathbf{u}_2+\cdots+\xi_n\mathbf{u}_n\|^2\\ &= \|\xi_1\mathbf{u}_1\|^2+\|\xi_2\mathbf{u}_2\|^2+\cdots+\|\xi_n\mathbf{u}_n\|^2\\ &= |\xi_1|^2+|\xi_2|^2+\cdots+|\xi_n|^2.\end{aligned}$$

5.4.16. Use the properties of an inner product to write

$$\begin{aligned}\left\|\mathbf{x}-\sum_{i=1}^k\xi_i\mathbf{u}_i\right\|^2 &= \left\langle\mathbf{x}-\sum_{i=1}^k\xi_i\mathbf{u}_i\,\middle|\,\mathbf{x}-\sum_{i=1}^k\xi_i\mathbf{u}_i\right\rangle\\ &= \langle\mathbf{x}|\mathbf{x}\rangle-2\sum_i|\xi_i|^2+\left\langle\sum_{i=1}^k\xi_i\mathbf{u}_i\,\middle|\,\sum_{i=1}^k\xi_i\mathbf{u}_i\right\rangle\\ &= \|\mathbf{x}\|^2-2\sum_i|\xi_i|^2+\left\|\sum_{i=1}^k\xi_i\mathbf{u}_i\right\|^2,\end{aligned}$$

and then invoke the Pythagorean theorem (Exercise 5.4.14) to conclude

$$\left\|\sum_{i=1}^k\xi_i\mathbf{u}_i\right\|^2 = \sum_i\|\xi_i\mathbf{u}_i\|^2 = \sum_i|\xi_i|^2.$$

Consequently,

$$0\le\left\|\mathbf{x}-\sum_{i=1}^k\xi_i\mathbf{u}_i\right\|^2 = \|\mathbf{x}\|^2-\sum_i|\xi_i|^2 \implies \sum_{i=1}^k|\xi_i|^2\le\|\mathbf{x}\|^2. \qquad (\ddagger)$$

If $\mathbf{x}\in span\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_k\}$, then the Fourier expansion of $\mathbf{x}$ with respect to the $\mathbf{u}_i$'s is $\mathbf{x}=\sum_{i=1}^k\xi_i\mathbf{u}_i$, and hence equality holds in $(\ddagger)$. Conversely, if equality holds in $(\ddagger)$, then $\mathbf{x}-\sum_{i=1}^k\xi_i\mathbf{u}_i=\mathbf{0}$.

5.4.17. Choose any unit vector $\mathbf{e}_i$ for $\mathbf{y}$. The angle between $\mathbf{e}$ and $\mathbf{e}_i$ approaches $\pi/2$ as $n \to \infty$, but $\mathbf{e}^T\mathbf{e}_i = 1$ for all n.

5.4.18. If $\mathbf{y}$ is negatively correlated to $\mathbf{x}$, then $\mathbf{z_x} = -\mathbf{z_y}$, but $\|\mathbf{z_x} - \mathbf{z_y}\|_2 = 2\sqrt{n}$ gives no indication of the fact that $\mathbf{z_x}$ and $\mathbf{z_y}$ are on the same line. Continuity therefore dictates that when $\mathbf{y} \approx \beta_0\mathbf{e} + \beta_1\mathbf{x}$ with $\beta_1 < 0$, then $\mathbf{z_x} \approx -\mathbf{z_y}$, but $\|\mathbf{z_x} - \mathbf{z_y}\|_2 \approx 2\sqrt{n}$ gives no hint that $\mathbf{z_x}$ and $\mathbf{z_y}$ are almost on the same line. If we want to use norms to gauge linear correlation, we should use

$$\min\left\{\|\mathbf{z_x} - \mathbf{z_y}\|_2\,,\ \|\mathbf{z_x} + \mathbf{z_y}\|_2\right\}.$$

5.4.19. (a) $\cos\theta = 1 \implies \langle\mathbf{x}|\mathbf{y}\rangle = \|\mathbf{x}\|\ \|\mathbf{y}\| > 0$, and the straightforward extension of Exercise 5.1.9 guarantees that

$$\mathbf{y} = \frac{\langle\mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|^2}\mathbf{x}, \quad \text{and clearly} \quad \frac{\langle\mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|^2} > 0.$$

Conversely, if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha > 0$, then $\langle\mathbf{x}|\mathbf{y}\rangle = \alpha\,\|\mathbf{x}\|^2 \implies \cos\theta = 1$.

(b) $\cos\theta = -1 \implies \langle\mathbf{x}|\mathbf{y}\rangle = -\,\|\mathbf{x}\|\ \|\mathbf{y}\| < 0$, so the generalized version of Exercise 5.1.9 guarantees that

$$\mathbf{y} = \frac{\langle\mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|^2}\mathbf{x}, \quad \text{and in this case} \quad \frac{\langle\mathbf{x}|\mathbf{y}\rangle}{\|\mathbf{x}\|^2} < 0.$$

Conversely, if $\mathbf{y} = \alpha\mathbf{x}$ for $\alpha < 0$, then $\langle\mathbf{x}|\mathbf{y}\rangle = \alpha\,\|\mathbf{x}\|^2$, so

$$\cos\theta = \frac{\alpha\,\|\mathbf{x}\|^2}{|\alpha|\ \|\mathbf{x}\|^2} = -1.$$

5.4.20. $F(t) = \sum_n^\infty (-1)^n \frac{2}{n}\sin nt$.

Solutions for exercises in section 5. 5

5.5.1. (a)

$$\mathbf{u}_1 = \frac{1}{2}\begin{pmatrix}1\\1\\1\\-1\end{pmatrix}, \quad \mathbf{u}_2 = \frac{1}{2\sqrt{3}}\begin{pmatrix}3\\-1\\-1\\1\end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{6}}\begin{pmatrix}0\\1\\1\\2\end{pmatrix}$$

(b) First verify this is an orthonormal set by showing $\mathbf{u}_i^T\mathbf{u}_j = \begin{cases}1 & \text{when } i = j,\\ 0 & \text{when } i \neq j.\end{cases}$

To show that the $\mathbf{x}_i$'s and the $\mathbf{u}_i$'s span the same space, place the $\mathbf{x}_i$'s as rows in a matrix $\mathbf{A}$, and place the $\mathbf{u}_i$'s as rows in a matrix $\mathbf{B}$, and then verify that $\mathbf{E_A} = \mathbf{E_B}$—recall Example 4.2.2.

(c) The result should be the same as in part (a).

5.5.2. First reduce $\mathbf{A}$ to $\mathbf{E_A}$ to determine a "regular" basis for each space.

$$R(\mathbf{A}) = span\left\{\begin{pmatrix}1\\2\\3\end{pmatrix}\right\} \qquad N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix}-2\\1\\0\end{pmatrix}, \begin{pmatrix}-3\\0\\1\end{pmatrix}\right\}$$

$$R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix}1\\-2\\3\\-1\end{pmatrix}\right\} \qquad N(\mathbf{A}) = span\left\{\begin{pmatrix}2\\1\\0\\0\end{pmatrix}, \begin{pmatrix}-3\\0\\1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\\1\end{pmatrix}\right\}$$

Now apply Gram–Schmidt to each of these.

$$R(\mathbf{A}) = span\left\{\frac{1}{\sqrt{14}}\begin{pmatrix}1\\2\\3\end{pmatrix}\right\} \quad N\left(\mathbf{A}^T\right) = span\left\{\frac{1}{\sqrt{5}}\begin{pmatrix}-2\\1\\0\end{pmatrix}, \frac{1}{\sqrt{70}}\begin{pmatrix}-3\\-6\\5\end{pmatrix}\right\}$$

$$R\left(\mathbf{A}^T\right) = span\left\{\frac{1}{\sqrt{15}}\begin{pmatrix}1\\-2\\3\\-1\end{pmatrix}\right\}$$

$$N(\mathbf{A}) = span\left\{\frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\\0\\0\end{pmatrix}, \frac{1}{\sqrt{70}}\begin{pmatrix}-3\\6\\5\\0\end{pmatrix}, \frac{1}{\sqrt{210}}\begin{pmatrix}1\\-2\\3\\14\end{pmatrix}\right\}$$

5.5.3.

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix}\mathrm{i}\\\mathrm{i}\\\mathrm{i}\end{pmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{6}}\begin{pmatrix}-2\mathrm{i}\\\mathrm{i}\\\mathrm{i}\end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\-\mathrm{i}\\\mathrm{i}\end{pmatrix}$$

5.5.4. Nothing! The resulting orthonormal set is the same as the original.

5.5.5. It breaks down at the first vector such that $\mathbf{x}_k \in span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$ because if

$$\mathbf{x}_k \in span\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\} = span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}\},$$

then the Fourier expansion of $\mathbf{x}_k$ with respect to $span\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{k-1}\}$ is

$$\mathbf{x}_k = \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \mathbf{u}_i,$$

and therefore

$$\mathbf{u}_k = \frac{\left(\mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \mathbf{u}_i\right)}{\left\|\left(\mathbf{x}_k - \sum_{i=1}^{k-1} \langle \mathbf{u}_i | \mathbf{x}_k \rangle \mathbf{u}_i\right)\right\|} = \frac{\mathbf{0}}{\|\mathbf{0}\|}$$

is not defined.

5.5.6. (a) The rectangular QR factors are

$$\mathbf{Q} = \begin{pmatrix} 1/\sqrt{3} & -1/\sqrt{3} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \\ 0 & 1/\sqrt{3} & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} \sqrt{3} & \sqrt{3} & -\sqrt{3} \\ 0 & \sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{6} \end{pmatrix}.$$

(b) Following Example 5.5.3, solve $\mathbf{R}\mathbf{x} = \mathbf{Q}^T\mathbf{b}$ to get $\mathbf{x} = \begin{pmatrix} 2/3 \\ 1/3 \\ 0 \end{pmatrix}$.

5.5.7. For $k = 1$, there is nothing to prove. For $k > 1$, assume that $\mathcal{O}_k$ is an orthonormal basis for $\mathcal{S}_k$. First establish that $\mathcal{O}_{k+1}$ must be an orthonormal set. Orthogonality follows because for each $j < k+1$,

$$\begin{aligned} \langle \mathbf{u}_j | \mathbf{u}_{k+1} \rangle &= \left\langle \mathbf{u}_j \middle| \frac{1}{\nu_{k+1}} \left(\mathbf{x}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right) \right\rangle \\ &= \frac{1}{\nu_{k+1}} \left(\langle \mathbf{u}_j | \mathbf{x}_{k+1} \rangle - \left\langle \mathbf{u}_j \middle| \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \right\rangle \right) \\ &= \frac{1}{\nu_{k+1}} \left(\langle \mathbf{u}_j | \mathbf{x}_{k+1} \rangle - \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \langle \mathbf{u}_j | \mathbf{u}_i \rangle \right) \\ &= \frac{1}{\nu_{k+1}} \left(\langle \mathbf{u}_j | \mathbf{x}_{k+1} \rangle - \langle \mathbf{u}_j | \mathbf{x}_{k+1} \rangle \right) = 0. \end{aligned}$$

This together with the fact that each $\mathbf{u}_i$ has unit norm means that $\mathcal{O}_{k+1}$ is an orthonormal set. Now assume $\mathcal{O}_k$ is a basis for $\mathcal{S}_k$, and prove that $\mathcal{O}_{k+1}$ is a basis for $\mathcal{S}_{k+1}$. If $\mathbf{x} \in \mathcal{S}_{k+1}$, then $\mathbf{x}$ can be written as a combination

$$\mathbf{x} = \sum_{i=1}^{k+1} \alpha_i \mathbf{x}_i = \left(\sum_{i=1}^{k} \alpha_i \mathbf{x}_i \right) + \alpha_{k+1} \mathbf{x}_{k+1},$$

where $\sum_{i=1}^{k} \alpha_i \mathbf{x}_i \in \mathcal{S}_k = span\,(\mathcal{O}_k) \subset span\,(\mathcal{O}_{k+1})$. Couple this together with the fact that

$$\mathbf{x}_{k+1} = \nu_{k+1} \mathbf{u}_{k+1} + \sum_{i=1}^{k} \langle \mathbf{u}_i | \mathbf{x}_{k+1} \rangle \mathbf{u}_i \in span\,(\mathcal{O}_{k+1})$$

to conclude that $\mathbf{x} \in span\,(\mathcal{O}_{k+1})$. Consequently, $\mathcal{O}_{k+1}$ spans $\mathcal{S}_{k+1}$, and therefore $\mathcal{O}_{k+1}$ is a basis for $\mathcal{S}_{k+1}$ because orthonormal sets are always linearly independent.

5.5.8. If $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$ are two rectangular QR factorizations, then (5.5.6) implies $\mathbf{A}^T\mathbf{A} = \mathbf{R}_1^T\mathbf{R}_1 = \mathbf{R}_2^T\mathbf{R}_2$. It follows from Example 3.10.7 that $\mathbf{A}^T\mathbf{A}$ is positive definite, and $\mathbf{R}_1 = \mathbf{R}_2$ because the Cholesky factorization of a positive definite matrix is unique. Therefore, $\mathbf{Q}_1 = \mathbf{A}\mathbf{R}_1^{-1} = \mathbf{A}\mathbf{R}_2^{-1} = \mathbf{Q}_2$.

5.5.9. (a) **Step 1:** $fl\,\|\mathbf{x}_1\| = 1$, so $\mathbf{u}_1 \leftarrow \mathbf{x}_1$.

Step 2: $\mathbf{u}_1^T\mathbf{x}_2 = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{x}_2 - \left(\mathbf{u}_1^T\mathbf{x}_2\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

Step 3: $\mathbf{u}_1^T\mathbf{x}_3 = 1$ and $\mathbf{u}_2^T\mathbf{x}_3 = 0$, so

$$\mathbf{u}_3 \leftarrow \mathbf{x}_3 - \left(\mathbf{u}_1^T\mathbf{x}_3\right)\mathbf{u}_1 - \left(\mathbf{u}_2^T\mathbf{x}_3\right)\mathbf{u}_2 = \begin{pmatrix} 0 \\ 10^{-3} \\ -10^{-3} \end{pmatrix} \text{ and } \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \begin{pmatrix} 0 \\ .709 \\ -.709 \end{pmatrix}.$$

Therefore, the result of the classical Gram–Schmidt algorithm using 3-digit arithmetic is

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ .709 \\ -.709 \end{pmatrix},$$

which is not very good because $\mathbf{u_2}$ and $\mathbf{u}_3$ are not even close to being orthogonal.

(b) **Step 1:** $fl\,\|\mathbf{x}_1\| = 1$, so

$$\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} \leftarrow \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}.$$

Step 2: $\mathbf{u}_1^T\mathbf{u}_2 = 1$ and $\mathbf{u}_1^T\mathbf{u}_3 = 1$, so

$$\mathbf{u}_2 \leftarrow \mathbf{u}_2 - \left(\mathbf{u}_1^T\mathbf{u}_2\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 0 \\ -10^{-3} \end{pmatrix}, \quad \mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_1^T\mathbf{u}_3\right)\mathbf{u}_1 = \begin{pmatrix} 0 \\ 10^{-3} \\ -10^{-3} \end{pmatrix},$$

and then

$$\mathbf{u}_2 \leftarrow \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

Step 3: $\mathbf{u}_2^T\mathbf{u}_3 = 10^{-3}$, so

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \left(\mathbf{u}_2^T\mathbf{u}_3\right)\mathbf{u}_2 = \begin{pmatrix} 0 \\ 10^{-3} \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_3 \leftarrow \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Thus the modified Gram–Schmidt algorithm produces

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 0 \\ 10^{-3} \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

which is as close to being an orthonormal set as one could reasonably hope to obtain by using 3-digit arithmetic.

5.5.10. Yes. In both cases r_{ij} is the (i,j)-entry in the upper-triangular matrix R in the QR factorization.

5.5.11. $p_0(x) = 1/\sqrt{2}, \quad p_1(x) = \sqrt{3/2}\,x, \quad p_2(x) = \sqrt{5/8}\,(3x^2 - 1)$

Solutions for exercises in section 5. 6

5.6.1. (a), (c), and (d).

5.6.2. Yes, because $\mathbf{U}^*\mathbf{U} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

5.6.3. (a) Eight: $\mathbf{D} = \begin{pmatrix} \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 \end{pmatrix}$ (b) 2^n : $\mathbf{D} = \begin{pmatrix} \pm 1 & 0 & \cdots & 0 \\ 0 & \pm 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pm 1 \end{pmatrix}$

(c) There are infinitely many because each diagonal entry can be any point on the unit circle in the complex plane—these matrices have the form given in part (d) of Exercise 5.6.1.

5.6.4. (a) When $\alpha^2 + \beta^2 = 1/2$. (b) When $\alpha^2 + \beta^2 = 1$.

5.6.5. (a) $(\mathbf{UV})^*(\mathbf{UV}) = \mathbf{V}^*\mathbf{U}^*\mathbf{UV} = \mathbf{V}^*\mathbf{V} = \mathbf{I}$.

(b) Consider $\mathbf{I} + (-\mathbf{I}) = \mathbf{0}$.

(c)

$$\begin{aligned} \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix}^* \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} &= \begin{pmatrix} \mathbf{U}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^* \end{pmatrix} \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{U}^*\mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^*\mathbf{V} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \end{aligned}$$

5.6.6. Recall from (3.7.8) or (4.2.10) that $(\mathbf{I}+\mathbf{A})^{-1}$ exists if and only if $N(\mathbf{I}+\mathbf{A}) = \mathbf{0}$, and write $\mathbf{x} \in N(\mathbf{I}+\mathbf{A}) \implies \mathbf{x} = -\mathbf{A}\mathbf{x} \implies \mathbf{x}^*\mathbf{x} = -\mathbf{x}^*\mathbf{A}\mathbf{x}$. But taking the conjugate transpose of both sides yields $\mathbf{x}^*\mathbf{x} = -\mathbf{x}^*\mathbf{A}^*\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x}$, so $\mathbf{x}^*\mathbf{x} = 0$, and thus $\mathbf{x} = \mathbf{0}$. Replacing $\mathbf{A}$ by $-\mathbf{A}$ in Exercise 3.7.6 gives $\mathbf{A}(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}\mathbf{A}$, so

$$\begin{aligned} (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} &= (\mathbf{I}+\mathbf{A})^{-1} - \mathbf{A}(\mathbf{I}+\mathbf{A})^{-1} \\ &= (\mathbf{I}+\mathbf{A})^{-1} - (\mathbf{I}+\mathbf{A})^{-1}\mathbf{A} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A}). \end{aligned}$$

These results together with the fact that $\mathbf{A}$ is skew hermitian produce

$$\begin{aligned}\mathbf{U}^*\mathbf{U} &= (\mathbf{I}+\mathbf{A})^{-1^*}(\mathbf{I}-\mathbf{A})^*(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1}\\ &= (\mathbf{I}+\mathbf{A})^{*^{-1}}(\mathbf{I}-\mathbf{A})^*(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1}\\ &= (\mathbf{I}-\mathbf{A})^{-1}(\mathbf{I}+\mathbf{A})(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = \mathbf{I}.\end{aligned}$$

5.6.7. (a) Yes—because if $\mathbf{R} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^*$, where $\|\mathbf{u}\| = 1$, then

$$\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix} = \mathbf{I} - 2\begin{pmatrix}\mathbf{0}\\ \mathbf{u}\end{pmatrix}\begin{pmatrix}\mathbf{0} & \mathbf{u}^*\end{pmatrix} \quad \text{and} \quad \left\|\begin{pmatrix}\mathbf{0}\\ \mathbf{u}\end{pmatrix}\right\| = 1.$$

(b) No—Suppose $\mathbf{R} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^*$ and $\mathbf{S} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^*$, where $\|\mathbf{u}\| = 1$ and $\|\mathbf{v}\| = 1$ so that

$$\begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{S}\end{pmatrix} = \mathbf{I} - 2\begin{pmatrix}\mathbf{u}\mathbf{u}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{v}\mathbf{v}^*\end{pmatrix}.$$

If we could find a vector $\mathbf{w}$ such that $\|\mathbf{w}\| = 1$ and

$$\begin{pmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{S}\end{pmatrix} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^*, \quad \text{then} \quad \mathbf{w}\mathbf{w}^* = \begin{pmatrix}\mathbf{u}\mathbf{u}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{v}\mathbf{v}^*\end{pmatrix}.$$

But this is impossible because (recall Example 3.9.3)

$$rank\,(\mathbf{w}\mathbf{w}^*) = 1 \quad \text{and} \quad rank\begin{pmatrix}\mathbf{u}\mathbf{u}^* & \mathbf{0}\\ \mathbf{0} & \mathbf{v}\mathbf{v}^*\end{pmatrix} = 2.$$

5.6.8. (a) $\mathbf{u}^*\mathbf{v} = (\mathbf{U}\mathbf{x})^*\mathbf{U}\mathbf{y} = \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{y} = \mathbf{x}^*\mathbf{y}$

(b) The fact that $\mathbf{P}$ is an isometry means $\|\mathbf{u}\| = \|\mathbf{x}\|$ and $\|\mathbf{v}\| = \|\mathbf{y}\|$. Use this together with part (a) and the definition of cosine given in (5.4.1) to obtain

$$\cos\theta_{\mathbf{u},\mathbf{v}} = \frac{\mathbf{u}^T\mathbf{v}}{\|\mathbf{u}\|\ \|\mathbf{v}\|} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\ \|\mathbf{y}\|} = \cos\theta_{\mathbf{x},\mathbf{y}}.$$

5.6.9. (a) Since $\mathbf{U}_{m\times r}$ has orthonormal columns, we have $\mathbf{U}^*\mathbf{U} = \mathbf{I}_r$ so that

$$\|\mathbf{U}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{x} = 1.$$

This together with $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2$—recall (5.2.10)—implies $\|\mathbf{V}\|_2 = 1$. For the Frobenius norm we have

$$\|\mathbf{U}\|_F = [trace\,(\mathbf{U}^*\mathbf{U})]^{1/2} = [trace\,(\mathbf{I})]^{1/2} = \sqrt{r}.$$

$$trace\,(\mathbf{A}\mathbf{B}) = trace\,(\mathbf{B}\mathbf{A}) \ \text{(Example 3.6.5) and } \mathbf{V}\mathbf{V}^* = \mathbf{I}_k \implies \|\mathbf{V}\|_F = \sqrt{k}.$$

(b) First show that $\|\mathbf{UA}\|_2 = \|\mathbf{A}\|_2$ by writing

$$\begin{aligned}\|\mathbf{UA}\|_2^2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{UAx}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}^*\mathbf{U}^*\mathbf{UAx} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}^*\mathbf{Ax} \\ &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2^2 = \|\mathbf{A}\|_2^2 .\end{aligned}$$

Now use this together with $\|\mathbf{A}\|_2 = \|\mathbf{A}^*\|_2$ to observe that

$$\|\mathbf{AV}\|_2 = \|\mathbf{V}^*\mathbf{A}^*\|_2 = \|\mathbf{A}^*\|_2 = \|\mathbf{A}\|_2 .$$

Therefore, $\|\mathbf{UAV}\|_2 = \|\mathbf{U}(\mathbf{AV})\|_2 = \|\mathbf{AV}\|_2 = \|\mathbf{A}\|_2$.

(c) Use $trace\,(\mathbf{AB}) = trace\,(\mathbf{BA})$ with $\mathbf{U}^*\mathbf{U} = \mathbf{I}_r$ and $\mathbf{VV}^* = \mathbf{I}_k$ to write

$$\begin{aligned}\|\mathbf{UAV}\|_F^2 &= trace\left((\mathbf{UAV})^*\mathbf{UAV}\right) = trace\,(\mathbf{V}^*\mathbf{A}^*\mathbf{U}^*\mathbf{UAV}) \\ &= trace\,(\mathbf{V}^*\mathbf{A}^*\mathbf{AV}) = trace\,(\mathbf{A}^*\mathbf{AVV}^*) \\ &= trace\,(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2 .\end{aligned}$$

5.6.10. Use (5.6.6) to compute the following quantities.

(a) $$\frac{\mathbf{vv}^T}{\mathbf{v}^T\mathbf{v}}\mathbf{u} = \left(\frac{\mathbf{v}^T\mathbf{u}}{\mathbf{v}^T\mathbf{v}}\right)\mathbf{v} = \frac{1}{6}\mathbf{v} = \frac{1}{6}\begin{pmatrix}1\\4\\0\\-1\end{pmatrix}$$

(b) $$\frac{\mathbf{uu}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{v} = \left(\frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u} = \frac{1}{5}\mathbf{u} = \frac{1}{5}\begin{pmatrix}-2\\1\\3\\-1\end{pmatrix}$$

(c) $$\left(\mathbf{I} - \frac{\mathbf{vv}^T}{\mathbf{v}^T\mathbf{v}}\right)\mathbf{u} = \mathbf{u} - \left(\frac{\mathbf{v}^T\mathbf{u}}{\mathbf{v}^T\mathbf{v}}\right)\mathbf{v} = \mathbf{u} - \frac{1}{6}\mathbf{v} = \frac{1}{6}\begin{pmatrix}-13\\2\\18\\-5\end{pmatrix}$$

(d) $$\left(\mathbf{I} - \frac{\mathbf{uu}^T}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{v} = \mathbf{v} - \left(\frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u} = \mathbf{v} - \frac{1}{5}\mathbf{u} = \frac{1}{5}\begin{pmatrix}7\\19\\-3\\-4\end{pmatrix}$$

5.6.11. (a) $N\,(\mathbf{Q}) \neq \{\mathbf{0}\}$ because $\mathbf{Qu} = \mathbf{0}$ and $\|\mathbf{u}\| = 1 \implies \mathbf{u} \neq \mathbf{0}$, so $\mathbf{Q}$ must be singular by (4.2.10).

(b) The result of Exercise 4.4.10 insures that $n-1 \leq rank\,(\mathbf{Q})$, and the result of part (a) says $rank\,(\mathbf{Q}) \leq n-1$, and therefore $rank\,(\mathbf{Q}) = n-1$.

5.6.12. Use (5.6.5) in conjunction with the CBS inequality given in (5.1.3) to write

$$\|\mathbf{p}\| = |\mathbf{u}^*\mathbf{x}| \leq \|\mathbf{u}\|\ \|\mathbf{x}\| = \|\mathbf{x}\| .$$

The fact that equality holds if and only if $\mathbf{x}$ is a scalar multiple of $\mathbf{u}$ follows from the result of Exercise 5.1.9.

5.6.13. (a) Set $\mathbf{u} = \mathbf{x} - \|\mathbf{x}\|\,\mathbf{e}_1 = -2/3\begin{pmatrix}1\\1\\1\end{pmatrix}$, and compute

$$\mathbf{R} = \mathbf{I} - \frac{2\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{3}\begin{pmatrix}1 & -2 & -2\\ -2 & 1 & -2\\ -2 & -2 & 1\end{pmatrix}.$$

(You could also use $\mathbf{u} = \mathbf{x} + \|\mathbf{x}\|\,\mathbf{e}_1$.)

(b) Verify that $\mathbf{R} = \mathbf{R}^T$, $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, and $\mathbf{R}^2 = \mathbf{I}$.

(c) The columns of the reflector $\mathbf{R}$ computed in part (a) do the job.

5.6.14. $\mathbf{R}\mathbf{x} = \mathbf{x} \implies 2\mathbf{u}\mathbf{u}^*\mathbf{x} = \mathbf{0} \implies \mathbf{u}^*\mathbf{x} = 0$ because $\mathbf{u} \neq \mathbf{0}$.

5.6.15. If $\mathbf{R}\mathbf{x} = \mathbf{y}$ in Figure 5.6.2, then the line segment between $\mathbf{x} - \mathbf{y}$ is parallel to the line determined by $\mathbf{u}$, so $\mathbf{x} - \mathbf{y}$ itself must be a scalar multiple of $\mathbf{u}$. If $\mathbf{x} - \mathbf{y} = \alpha\mathbf{u}$, then

$$\mathbf{u} = \frac{\mathbf{x} - \mathbf{y}}{\alpha} = \frac{\mathbf{x} - \mathbf{y}}{\|\mathbf{x} - \mathbf{y}\|}.$$

It is straightforward to verify that this choice of $\mathbf{u}$ produces the desired reflector.

5.6.16. You can verify by direct multiplication that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, but you can also recognize that $\mathbf{P}$ and $\mathbf{U}$ are elementary reflectors that come from Example 5.6.3 in the sense that

$$\mathbf{P} = \mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}, \quad \text{where} \quad \mathbf{u} = \mathbf{x} - \mathbf{e}_1 = \begin{pmatrix}x_1 - 1\\ \tilde{\mathbf{x}}\end{pmatrix}$$

and

$$\mathbf{U} = \mu\left(\mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^*}{\mathbf{u}^*\mathbf{u}}\right), \quad \text{where} \quad \mathbf{u} = \mathbf{x} - \mu\mathbf{e}_1 = \begin{pmatrix}x_1 - \mu\\ \tilde{\mathbf{x}}\end{pmatrix}.$$

5.6.17. The final result is

$$\mathbf{v}_3 = \begin{pmatrix}-\sqrt{2}/2\\ \sqrt{6}/2\\ 1\end{pmatrix}$$

and

$$\mathbf{Q} = \mathbf{P}_z(\pi/6)\mathbf{P}_y(-\pi/2)\mathbf{P}_x(\pi/4) = \frac{1}{4}\begin{pmatrix}0 & -\sqrt{6}-\sqrt{2} & -\sqrt{6}+\sqrt{2}\\ 0 & \sqrt{6}-\sqrt{2} & -\sqrt{6}-\sqrt{2}\\ 4 & 0 & 0\end{pmatrix}.$$

5.6.18. It matters because the rotation matrices given on p. 328 generally do not commute with each other (this is easily verified by direct multiplication). For example, this means that it is generally the case that

$$\mathbf{P}_y(\phi)\mathbf{P}_x(\theta)\mathbf{v} \neq \mathbf{P}_x(\theta)\mathbf{P}_y(\phi)\mathbf{v}.$$

5.6.19. As pointed out in Example 5.6.2, $\mathbf{u}^\perp = (\mathbf{u}/\|\mathbf{u}\|)^\perp$, so we can assume without any loss of generality that $\mathbf{u}$ has unit norm. We also know that any vector of unit norm can be extended to an orthonormal basis for $\mathcal{C}^n$—Examples 5.6.3 and 5.6.6 provide two possible ways to accomplish this. Let $\{\mathbf{u}, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$ be such an orthonormal basis for $\mathcal{C}^n$.

Claim: $span\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\} = \mathbf{u}^\perp$.

Proof. $\mathbf{x} \in span\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\} \implies \mathbf{x} = \sum_i \alpha_i \mathbf{v}_i \implies \mathbf{u}^*\mathbf{x} = \sum_i \alpha_i \mathbf{u}^*\mathbf{v}_i = 0 \implies \mathbf{x} \in \mathbf{u}^\perp$, and thus $span\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\} \subseteq \mathbf{u}^\perp$. To establish the reverse inclusion, write $\mathbf{x} = \alpha_0\mathbf{u} + \sum_i \alpha_i \mathbf{v}_i$, and then note that $\mathbf{x} \perp \mathbf{u} \implies 0 = \mathbf{u}^*\mathbf{x} = \alpha_0 \implies \mathbf{x} \in span\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$, and hence $\implies \mathbf{u}^\perp \subseteq span\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$.

Consequently, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n-1}\}$ is a basis for $\mathbf{u}^\perp$ because it is a spanning set that is linearly independent—recall (4.3.14)—and thus $\dim \mathbf{u}^\perp = n - 1$.

5.6.20. The relationship between the matrices in (5.6.6) and (5.6.7) on p. 324 suggests that if $\mathbf{P}$ is a projector, then $\mathbf{A} = \mathbf{I} - 2\mathbf{P}$ is an involution—and indeed this is true because $\mathbf{A}^2 = (\mathbf{I} - 2\mathbf{P})^2 = \mathbf{I} - 4\mathbf{P} + 4\mathbf{P}^2 = \mathbf{I}$. Similarly, if $\mathbf{A}$ is an involution, then $\mathbf{P} = (\mathbf{I} - \mathbf{A})/2$ is easily verified to be a projector. Thus each projector uniquely defines an involution, and vice versa.

5.6.21. The outside of the face is visible from the perspective indicated in Figure 5.6.6 if and only if the angle θ between $\mathbf{n}$ and the positive x-axis is between $-90°$ and $+90°$. This is equivalent to saying that the cosine between $\mathbf{n}$ and $\mathbf{e}_1$ is positive, so the desired conclusion follows from the fact that

$$\cos\theta > 0 \iff \frac{\mathbf{n}^T\mathbf{e}_1}{\|\mathbf{n}\|\,\|\mathbf{e}_1\|} > 0 \iff \mathbf{n}^T\mathbf{e}_1 > 0 \iff n_1 > 0.$$

Solutions for exercises in section 5. 7

5.7.1. (a) Householder reduction produces

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix} \begin{pmatrix} 1/3 & -2/3 & 2/3 \\ -2/3 & 1/3 & 2/3 \\ 2/3 & 2/3 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & 19 & -34 \\ -2 & -5 & 20 \\ 2 & 8 & 37 \end{pmatrix}$$
$$= \begin{pmatrix} 3 & 15 & 0 \\ 0 & 15 & -30 \\ 0 & 0 & 45 \end{pmatrix} = \mathbf{R},$$

so

$$\mathbf{Q} = (\mathbf{R}_2\mathbf{R}_1)^T = \begin{pmatrix} 1/3 & 14/15 & -2/15 \\ -2/3 & 1/3 & 2/3 \\ 2/3 & -2/15 & 11/15 \end{pmatrix}.$$

(b) Givens reduction produces $\mathbf{P}_{23}\mathbf{P}_{13}\mathbf{P}_{12}\mathbf{A} = \mathbf{R}$, where

$$\mathbf{P}_{12} = \begin{pmatrix} 1/\sqrt{5} & -2/\sqrt{5} & 0 \\ 2/\sqrt{5} & 1/\sqrt{5} & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbf{P}_{13} = \begin{pmatrix} \sqrt{5}/3 & 0 & 2/3 \\ 0 & 1 & 0 \\ -2/3 & 0 & \sqrt{5}/3 \end{pmatrix}$$

$$\mathbf{P}_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 11/5\sqrt{5} & -2/5\sqrt{5} \\ 0 & 2/5\sqrt{5} & 11/5\sqrt{5} \end{pmatrix}$$

5.7.2. Since $\mathbf{P}$ is an orthogonal matrix, so is $\mathbf{P}^T$, and hence the columns of $\mathbf{X}$ are an orthonormal set. By writing

$$\mathbf{A} = \mathbf{P}^T\mathbf{T} = [\mathbf{X} \,|\, \mathbf{Y}] \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} = \mathbf{XR},$$

and by using the fact that $rank\,(\mathbf{A}) = n \implies rank\,(\mathbf{R}) = n$, it follows that $R\,(\mathbf{A}) = R\,(\mathbf{XR}) = R\,(\mathbf{X})$—recall Exercise 4.5.12. Since every orthonormal set is linearly independent, the columns of $\mathbf{X}$ are a linearly independent spanning set for $R\,(\mathbf{A})$, and thus the columns of $\mathbf{X}$ are an orthonormal basis for $R\,(\mathbf{A})$. Notice that when the diagonal entries of $\mathbf{R}$ are positive, $\mathbf{A} = \mathbf{XR}$ is the "rectangular" QR factorization for $\mathbf{A}$ introduced on p. 311, and the columns of $\mathbf{X}$ are the same columns as those produced by the Gram–Schmidt procedure.

5.7.3. According to (5.7.1), set $\mathbf{u} = \mathbf{A}_{*1} - \|\mathbf{A}_{*1}\|\,\mathbf{e}_1 = \begin{pmatrix} -1 \\ 2 \\ -2 \\ 1 \end{pmatrix}$, so

$$\mathbf{R}_1 = \mathbf{I} - 2\frac{\mathbf{uu}^*}{\mathbf{u}^*\mathbf{u}} = \frac{1}{5}\begin{pmatrix} 4 & 2 & -2 & 1 \\ 2 & 1 & 4 & -2 \\ -2 & 4 & 1 & 2 \\ 1 & -2 & 2 & 4 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_1\mathbf{A} = \begin{pmatrix} 5 & -15 & 5 \\ 0 & 10 & -5 \\ 0 & -10 & 2 \\ 0 & 5 & 14 \end{pmatrix}.$$

Next use $\mathbf{u} = \begin{pmatrix} 10 \\ -10 \\ 5 \end{pmatrix} - \begin{pmatrix} 15 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -5 \\ -10 \\ 5 \end{pmatrix}$ to build

$$\hat{\mathbf{R}}_2 = \mathbf{I} - 2\frac{\mathbf{uu}^*}{\mathbf{u}^*\mathbf{u}} = \frac{1}{3}\begin{pmatrix} 2 & -2 & 1 \\ -2 & -1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_2 = \frac{1}{3}\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & -2 & 1 \\ 0 & -2 & -1 & 2 \\ 0 & 1 & 2 & 2 \end{pmatrix},$$

so

$$\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} 5 & -15 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 12 \\ 0 & 0 & 9 \end{pmatrix}.$$

Finally, with $\mathbf{u} = \begin{pmatrix} 12 \\ 9 \end{pmatrix} - \begin{pmatrix} 15 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 9 \end{pmatrix}$, build

$$\hat{\mathbf{R}}_3 = \frac{1}{5}\begin{pmatrix} 4 & 3 \\ 3 & -4 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_3 = \frac{1}{5}\begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 4 & 3 \\ 0 & 0 & 3 & -4 \end{pmatrix},$$

so that

$$\mathbf{R}_3\mathbf{R}_2\mathbf{R}_1\mathbf{A} = \begin{pmatrix} 5 & -15 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, $\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}$, where

$$\mathbf{P} = \mathbf{R}_3\mathbf{R}_2\mathbf{R}_1 = \frac{1}{15}\begin{pmatrix} 12 & 6 & -6 & 3 \\ 9 & -8 & 8 & -4 \\ 0 & -5 & 2 & 14 \\ 0 & -10 & -11 & -2 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} 5 & -15 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \end{pmatrix}.$$

The result of Exercise 5.7.2 insures that the first three columns in

$$\mathbf{P}^T = \mathbf{R}_1\mathbf{R}_2\mathbf{R}_3 = \frac{1}{15}\begin{pmatrix} 12 & 9 & 0 & 0 \\ 6 & -8 & -5 & -10 \\ -6 & 8 & 2 & -11 \\ 3 & -4 & 14 & -2 \end{pmatrix}$$

are an orthonormal basis for $R(\mathbf{A})$. Since the diagonal entries of $\mathbf{R}$ are positive,

$$\frac{1}{15}\begin{pmatrix} 12 & 9 & 0 \\ 6 & -8 & -5 \\ -6 & 8 & 2 \\ 3 & -4 & 14 \end{pmatrix}\begin{pmatrix} 5 & -15 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \end{pmatrix} = \mathbf{A}$$

is the "rectangular" QR factorization for $\mathbf{A}$ discussed on p. 311.

5.7.4. If $\mathbf{A}$ has full column rank, and if $\mathbf{P}$ is an orthogonal matrix such that

$$\mathbf{PA} = \mathbf{T} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \mathbf{Pb} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix},$$

where $\mathbf{R}$ is an upper-triangular matrix, then the results of Example 5.7.3 insure that the least squares solution of $\mathbf{Ax} = \mathbf{b}$ can be obtained by solving the triangular system $\mathbf{Rx} = \mathbf{c}$. The matrices $\mathbf{P}$ and $\mathbf{R}$ were computed in Exercise 5.7.3, so the least squares solution of $\mathbf{Ax} = \mathbf{b}$ is the solution to

$$\begin{pmatrix} 5 & -15 & 5 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 33 \end{pmatrix} \implies \mathbf{x} = \frac{1}{5}\begin{pmatrix} -4 \\ 1 \\ 11 \end{pmatrix}.$$

5.7.5. $\|\mathbf{A}\|_F = \|\mathbf{QR}\|_F = \|\mathbf{R}\|_F$ because orthogonal matrices are norm preserving transformations—recall Exercise 5.6.9.

5.7.6. Follow the procedure outlined in Example 5.7.4 to compute the reflector

$$\hat{\mathbf{R}} = \begin{pmatrix} -3/5 & 4/5 \\ 4/5 & 3/5 \end{pmatrix}, \quad \text{and then set} \quad \mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -3/5 & 4/5 \\ 0 & 4/5 & 3/5 \end{pmatrix}.$$

Since $\mathbf{A}$ is 3×3, there is only one step, so $\mathbf{P} = \mathbf{R}$ and

$$\mathbf{P}^T\mathbf{AP} = \mathbf{H} = \begin{pmatrix} -2 & -5 & 0 \\ -5 & -41 & 38 \\ 0 & 38 & 41 \end{pmatrix}.$$

5.7.7. First argue that the product of an upper-Hessenberg matrix with an upper-triangular matrix must be upper Hessenberg—regardless of which side the triangular factor appears. This implies that $\mathbf{Q}$ is upper Hessenberg because $\mathbf{Q} = \mathbf{HR}^{-1}$ and $\mathbf{R}^{-1}$ is upper triangular—recall Exercise 3.7.4. This in turn means that $\mathbf{RQ}$ must be upper Hessenberg.

5.7.8. From the structure of the matrices in Example 5.7.5, it can be seen that $\mathbf{P}_{12}$ requires $4n$ multiplications, $\mathbf{P}_{23}$ requires $4(n-1)$ multiplications, etc. Use the formula $1 + 2 + \cdots + n = n(n+1)/2$ to obtain the total as

$$4[n + (n-1) + (n-2) + \cdots + 2] = 4\left(\frac{n^2+n}{2} - 1\right) \approx 2n^2.$$

Solutions for exercises in section 5. 8

5.8.1. (a) $\begin{pmatrix} 4 \\ 13 \\ 28 \\ 27 \\ 18 \\ 0 \end{pmatrix}$ (b) $\begin{pmatrix} -1 \\ 0 \\ 2 \\ 0 \\ -1 \\ 0 \end{pmatrix}$ (c) $\begin{pmatrix} \alpha_0 \\ \alpha_0 + \alpha_1 \\ \alpha_0 + \alpha_1 + \alpha_2 \\ \alpha_1 + \alpha_2 \\ \alpha_2 \\ 0 \end{pmatrix}$

5.8.2. The answer to both parts is $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 4 \end{pmatrix}$.

5.8.3. $\mathbf{F}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, $\mathbf{D}_2 = \begin{pmatrix} 1 & 0 \\ 0 & -\mathrm{i} \end{pmatrix}$, and

$$\mathbf{F}_4\mathbf{P}_4^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix} \mathbf{P}_4^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -\mathrm{i} & \mathrm{i} \\ 1 & 1 & -1 & -1 \\ 1 & -1 & \mathrm{i} & -\mathrm{i} \end{pmatrix}$$

$$= \left(\begin{array}{cc|c} 1 & 1 & \begin{pmatrix} 1 & 0 \\ 0 & -\mathrm{i} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ 1 & -1 & \\ \hline 1 & 1 & -\begin{pmatrix} 1 & 0 \\ 0 & -\mathrm{i} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\ 1 & -1 & \end{array}\right) = \begin{pmatrix} \mathbf{F}_2 & \mathbf{D}_2\mathbf{F}_2 \\ \mathbf{F}_2 & -\mathbf{D}_2\mathbf{F}_2 \end{pmatrix}.$$

5.8.4. (a) $\mathbf{a} \odot \mathbf{b} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_0\beta_1 + \alpha_1\beta_0 \\ \alpha_1\beta_1 \\ 0 \end{pmatrix}$

$$\mathbf{F}_4(\mathbf{a} \odot \mathbf{b}) = \begin{pmatrix} \alpha_0\beta_0 + \alpha_0\beta_1 + \alpha_1\beta_0 + \alpha_1\beta_1 \\ \alpha_0\beta_0 - \mathrm{i}\alpha_0\beta_1 - \mathrm{i}\alpha_1\beta_0 - \alpha_1\beta_1 \\ \alpha_0\beta_0 - \alpha_0\beta_1 - \alpha_1\beta_0 + \alpha_1\beta_1 \\ \alpha_0\beta_0 + \mathrm{i}\alpha_0\beta_1 + \mathrm{i}\alpha_1\beta_0 - \alpha_1\beta_1 \end{pmatrix} = (\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})$$

(b) $\mathbf{F}_4^{-1}\big[(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})\big] = \mathbf{a} \odot \mathbf{b}$

5.8.5. $p(x)q(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3$, where

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \mathbf{F}_4^{-1}\big[(\mathbf{F}_4\hat{\mathbf{a}}) \times (\mathbf{F}_4\hat{\mathbf{b}})\big]$$

$$= \mathbf{F}_4^{-1}\left[\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix}\begin{pmatrix} -3 \\ 2 \\ 0 \\ 0 \end{pmatrix} \times \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix}\begin{pmatrix} -4 \\ 3 \\ 0 \\ 0 \end{pmatrix}\right]$$

$$= \mathbf{F}_4^{-1}\left[\begin{pmatrix} -1 \\ -3-2\mathrm{i} \\ -5 \\ -3+2\mathrm{i} \end{pmatrix} \times \begin{pmatrix} -1 \\ -4-3\mathrm{i} \\ -7 \\ -4+3\mathrm{i} \end{pmatrix}\right]$$

$$= \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}\begin{pmatrix} 1 \\ 6+17\mathrm{i} \\ 35 \\ 6-17\mathrm{i} \end{pmatrix} = \begin{pmatrix} 12 \\ -17 \\ 6 \\ 0 \end{pmatrix}.$$

5.8.6. (a) $\begin{pmatrix} 3 \\ 4 \end{pmatrix} \odot \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 10 \\ 8 \\ 0 \end{pmatrix}$, so

$$\begin{aligned} 43_{10} \times 21_{10} &= (8 \times 10^2) + (10 \times 10^1) + (3 \times 10^0) \\ &= (9 \times 10^2) + (0 \times 10^1) + (3 \times 10^0) = 903. \end{aligned}$$

(b) $$\begin{pmatrix}1\\2\\3\end{pmatrix} \odot \begin{pmatrix}6\\0\\1\end{pmatrix} = \begin{pmatrix}3\\2\\19\\12\\6\\0\end{pmatrix}, \text{ so}$$

$$123_8 \times 601_8 = (6 \times 8^4) + (12 \times 8^3) + (19 \times 8^2) + (2 \times 8^1) + (3 \times 8^0).$$

Since

$$\begin{aligned} 12 = 8 + 4 \quad &\Longrightarrow \quad 12 \times 8^3 = (8+4) \times 8^3 = 8^4 + (4 \times 8^3) \\ 19 = (2 \times 8) + 3 \quad &\Longrightarrow \quad 19 \times 8^2 = (2 \times 8^3) + (3 \times 8^2), \end{aligned}$$

we have that

$$123_8 \times 601_8 = (7 \times 8^4) + (6 \times 8^3) + (3 \times 8^2) + (2 \times 8^1) + (3 \times 8^0) = 76323_8.$$

(c) $$\begin{pmatrix}1\\0\\1\\0\end{pmatrix} \odot \begin{pmatrix}1\\1\\0\\1\end{pmatrix} = \begin{pmatrix}0\\1\\0\\2\\1\\1\\1\\0\end{pmatrix}, \text{ so}$$

$$1010_2 \times 1101_2 = (1\times 2^6) + (1\times 2^5) + (1\times 2^4) + (2\times 2^3) + (0\times 2^2) + (1\times 2^1) + (0\times 2^0).$$

Substituting $2 \times 2^3 = 1 \times 2^4$ in this expression and simplifying yields

$$\begin{aligned} 1010_2 \times 1101_2 &= (1 \times 2^7) + (0 \times 2^6) + (0 \times 2^5) + (0 \times 2^4) \\ &\quad + (0 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (0 \times 2^0) \\ &= 10000010_2. \end{aligned}$$

5.8.7. (a) The number of multiplications required by the definition is

$$\begin{aligned} & 1 + 2 + \cdots + (n-1) + n + (n-1) + \cdots + 2 + 1 \\ = \; & 2\Big(1 + 2 + \cdots + (n-1)\Big) + n \\ = \; & (n-1)n + n = n^2. \end{aligned}$$

(b) In the formula $\mathbf{a}_{n\times 1} \odot \mathbf{b}_{n\times 1} = \mathbf{F}_{2n}^{-1}\big[(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})\big]$, using the FFT to compute $\mathbf{F}_{2n}\hat{\mathbf{a}}$ and $\mathbf{F}_{2n}\hat{\mathbf{b}}$ requires $(2n/2)\log_2 2n = n(1 + \log_2 n)$ multiplications for each term, and an additional $2n$ multiplications are needed to form the product $(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})$. Using the FFT in conjunction with the procedure

described in Example 5.8.2 to apply $\mathbf{F}^{-1}$ to $(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})$ requires another $(2n/2)\log_2 2n = n(1+\log_2 n)$ multiplications to compute $\overline{\mathbf{F}_{2n}\overline{\mathbf{x}}}$ followed by $2n$ more multiplications to produce $(1/2n)\overline{\mathbf{F}_{2n}\overline{\mathbf{x}}} = \mathbf{F}_{2n}^{-1}\mathbf{x}$. Therefore, the total count is $3n(1+\log_2 n) + 4n = 3n\log_2 n + 7n$.

5.8.8. Recognize that $\mathbf{y}$ is of the form

$$\mathbf{y} = 1(\mathbf{e}_2+\mathbf{e}_6) + 4(\mathbf{e}_3+\mathbf{e}_5) + 5\mathrm{i}(-\mathbf{e}_1+\mathbf{e}_7) + 3\mathrm{i}(-\mathbf{e}_2+\mathbf{e}_6).$$

The real part says that there are two cosines—one with amplitude 1 and frequency 2, and the other with amplitude 4 and frequency 3. The imaginary part says there are two sines—one with amplitude 5 and frequency 1, and the other with amplitude 3 and frequency 2. Therefore,

$$x(\tau) = \cos 4\pi\tau + 4\cos 6\pi\tau + 5\sin 2\pi\tau + 3\sin 4\pi\tau.$$

5.8.9. Use (5.8.12) to write $\mathbf{a}\odot\mathbf{b} = \mathbf{F}^{-1}\big[(\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}})\big] = \mathbf{F}^{-1}\big[(\mathbf{F}\hat{\mathbf{b}}) \times (\mathbf{F}\hat{\mathbf{a}})\big] = \mathbf{a}\odot\mathbf{b}$.

5.8.10. This is a special case of the result given in Example 4.3.5. The Fourier matrix $\mathbf{F}_n$ is a special case of the Vandermonde matrix—simply let x_k's that define the Vandermonde matrix be the nth roots of unity.

5.8.11. The result of Exercise 5.8.10 implies that if

$$\hat{\mathbf{a}} = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1} \quad \text{and} \quad \hat{\mathbf{b}} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{n-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{2n\times 1},$$

then $\mathbf{F}_{2n}\hat{\mathbf{a}} = \mathbf{p}$ and $\mathbf{F}_{2n}\hat{\mathbf{b}} = \mathbf{q}$, and we know from (5.8.11) that the γ_k's are given by $\gamma_k = [\mathbf{a}\odot\mathbf{b}]_k$. Therefore, the convolution theorem guarantees

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \end{pmatrix} = \mathbf{a}\odot\mathbf{b} = \mathbf{F}_{2n}^{-1}\big[(\mathbf{F}_{2n}\hat{\mathbf{a}}) \times (\mathbf{F}_{2n}\hat{\mathbf{b}})\big] = \mathbf{F}_{2n}^{-1}\big[\mathbf{p}\times\mathbf{q}\big] = \mathbf{F}_{2n}^{-1}\begin{pmatrix} p(1)q(1) \\ p(\xi)q(\xi) \\ p(\xi^2)q(\xi^2) \\ \vdots \end{pmatrix}.$$

5.8.12. (a) This follows from the observation that $\mathbf{Q}^k$ has 1's on the kth subdiagonal and 1's on the $(n-k)^{th}$ superdiagonal. For example, if $n=8$, then

$$\mathbf{Q}^3 = \begin{pmatrix} 0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&1 \\ 1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0&0 \end{pmatrix}.$$

(b) If the rows of $\mathbf{F}$ are indexed from 0 to $n-1$, then they satisfy the relationships $\mathbf{F}_{k*}\mathbf{Q} = \xi^k \mathbf{F}_{k*}$ for each k (verifying this for $n = 4$ will indicate why it is true in general). This means that $\mathbf{FQ} = \mathbf{DF}$, which in turn implies $\mathbf{FQF}^{-1} = \mathbf{D}$.

(c) Couple parts (a) and (b) with $\mathbf{FQ}^k\mathbf{F}^{-1} = (\mathbf{FQF}^{-1})^k = \mathbf{D}^k$ to write

$$\begin{aligned}
\mathbf{FCF}^{-1} &= \mathbf{F}p(\mathbf{Q})\mathbf{F}^{-1} \\
&= \mathbf{F}(c_0\mathbf{I} + c_1\mathbf{Q} + \cdots + c_{n-1}\mathbf{Q}^{n-1})\mathbf{F}^{-1} \\
&= c_0\mathbf{I} + c_1\mathbf{FQF}^{-1} + \cdots + c_{n-1}\mathbf{FQ}^{n-1}\mathbf{F}^{-1} \\
&= c_0\mathbf{I} + c_1\mathbf{D} + \cdots + c_{n-1}\mathbf{D}^{n-1} \\
&= \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix}.
\end{aligned}$$

(d) $\mathbf{FC}_1\mathbf{F}^{-1} = \mathbf{D}_1$ and $\mathbf{FC}_2\mathbf{F}^{-1} = \mathbf{D}_2$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices, and therefore

$$\begin{aligned}
\mathbf{C}_1\mathbf{C}_2 &= \mathbf{F}^{-1}\mathbf{D}_1\mathbf{FF}^{-1}\mathbf{D}_2\mathbf{F} = \mathbf{F}^{-1}\mathbf{D}_1\mathbf{D}_2\mathbf{F} = \mathbf{F}^{-1}\mathbf{D}_2\mathbf{D}_1\mathbf{F} = \mathbf{F}^{-1}\mathbf{D}_2\mathbf{FF}^{-1}\mathbf{D}_1\mathbf{F} \\
&= \mathbf{C}_2\mathbf{C}_1.
\end{aligned}$$

5.8.13. (a) According to Exercise 5.8.12,

$$\mathbf{C} = \begin{pmatrix} \sigma_0 & \sigma_{n-1} & \cdots & \sigma_1 \\ \sigma_1 & \sigma_0 & \cdots & \sigma_2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n-1} & \sigma_{n-2} & \cdots & \sigma_0 \end{pmatrix} = \mathbf{F}^{-1} \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{n-1}) \end{pmatrix} \mathbf{F} = \mathbf{F}^{-1}\mathbf{DF}$$

in which $p(x) = \sigma_0 + \sigma_1 x + \cdots + \sigma_{n-1}x^{n-1}$. Therefore, $\mathbf{x} = \mathbf{C}^{-1}\mathbf{b} = \mathbf{F}^{-1}\mathbf{D}^{-1}\mathbf{Fb}$, so we can execute the following computations.

(i) $\begin{pmatrix} p(0) \\ p(\xi) \\ \vdots \\ p(\xi^{n-1}) \end{pmatrix} \longleftarrow \mathbf{F} \begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \vdots \\ \sigma_{n-1} \end{pmatrix}$ using the FFT

(ii) $\mathbf{x} \longleftarrow \mathbf{Fb}$ using the FFT

(iii) $x_k \longleftarrow x_k/p(\xi^k)$ for $k = 0, 1, \ldots, n-1$

(iv) $\mathbf{x} \longleftarrow \mathbf{F}^{-1}\mathbf{x}$ using the FFT as described in Example 5.8.2

(b) Use the same techniques described in part (a) to compute the kth column of $\mathbf{C}^{-1}$ from the formula

$$\begin{aligned}[\mathbf{C}^{-1}]_{*k} &= \mathbf{C}^{-1}\mathbf{e}_k = \mathbf{F}^{-1}\mathbf{D}^{-1}\mathbf{F}\mathbf{e}_k \\ &= \mathbf{F}^{-1}\left(\mathbf{D}^{-1}[\mathbf{F}]_{*k}\right) \\ &= \mathbf{F}^{-1}\begin{pmatrix} 1/p(1) \\ \xi^k/p(\xi) \\ \xi^{2k}/p(\xi^2) \\ \vdots \\ \xi^{n-k}/p(\xi^{n-1}) \end{pmatrix}.\end{aligned}$$

(c) The kth column of $\mathbf{P} = \mathbf{C}_1\mathbf{C}_2$ is given by

$$\mathbf{P}_{*k} = \mathbf{P}\mathbf{e}_k = \mathbf{F}^{-1}\mathbf{D}_1\mathbf{F}\mathbf{F}^{-1}\mathbf{D}_2\mathbf{F}\mathbf{e}_k = \mathbf{F}^{-1}\Big(\mathbf{D}_1\mathbf{D}_2[\mathbf{F}]_{*k}\Big).$$

If $(\,\sigma_0 \quad \sigma_1 \quad \cdots \quad \sigma_{n-1}\,)$ and $(\,\eta_0 \quad \eta_1 \quad \cdots \quad \eta_{n-1}\,)$ are the first rows in $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively, and if $p(x) = \sum_{k=0}^{n-1}\sigma_k x^k$ and $q(x) = \sum_{k=0}^{n-1}\eta_k x^k$, then first compute

$$\mathbf{p} = \begin{pmatrix} p(0) \\ p(\xi) \\ \vdots \\ p(\xi^{n-1}) \end{pmatrix} \longleftarrow \mathbf{F}\begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \vdots \\ \sigma_{n-1} \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} q(0) \\ q(\xi) \\ \vdots \\ q(\xi^{n-1}) \end{pmatrix} \longleftarrow \mathbf{F}\begin{pmatrix} \eta_0 \\ \eta_1 \\ \vdots \\ \eta_{n-1} \end{pmatrix}.$$

The kth column of the product can now be obtained from

$$\mathbf{P}_{*k} \longleftarrow \mathbf{F}^{-1}\Big(\mathbf{p} \times \mathbf{q} \times \mathbf{F}_{*k}\Big) \quad \text{for} \quad k = 0, 1, \ldots, n-1.$$

5.8.14. (a) For $n = 3$ we have

$$\mathbf{C}\hat{\mathbf{b}} = \begin{pmatrix} \alpha_0 & 0 & 0 & 0 & \alpha_2 & \alpha_1 \\ \alpha_1 & \alpha_0 & 0 & 0 & 0 & \alpha_2 \\ \alpha_2 & \alpha_1 & \alpha_0 & 0 & 0 & 0 \\ 0 & \alpha_2 & \alpha_1 & \alpha_0 & 0 & 0 \\ 0 & 0 & \alpha_2 & \alpha_1 & \alpha_0 & 0 \\ 0 & 0 & 0 & \alpha_2 & \alpha_1 & \alpha_0 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha_0\beta_0 \\ \alpha_1\beta_0 + \alpha_0\beta_1 \\ \alpha_2\beta_0 + \alpha_1\beta_1 + \alpha_0\beta_2 \\ \alpha_2\beta_1 + \alpha_1\beta_2 \\ \alpha_2\beta_2 \\ 0 \end{pmatrix}.$$

Use this as a model to write the expression for $\mathbf{C}\hat{\mathbf{b}}$, where n is arbitrary.

(b) We know from part (c) of Exercise 5.8.12 that if $\mathbf{F}$ is the Fourier matrix of order $2n$, then $\mathbf{F}\mathbf{C}\mathbf{F}^{-1} = \mathbf{D}$, where

$$\mathbf{D} = \begin{pmatrix} p(1) & 0 & \cdots & 0 \\ 0 & p(\xi) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(\xi^{2n-1}) \end{pmatrix} \quad (\text{the } \xi^k\text{'s are the } 2n^{th} \text{ roots of unity})$$

in which $p(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_{n-1}x^{n-1}$. Therefore, from part (a),

$$\mathbf{F}(\mathbf{a} \odot \mathbf{b}) = \mathbf{F}\mathbf{C}\hat{\mathbf{b}} = \mathbf{F}\mathbf{C}\mathbf{F}^{-1}\mathbf{F}\hat{\mathbf{b}} = \mathbf{D}\mathbf{F}\hat{\mathbf{b}}.$$

According to Exercise 5.8.10, we also know that

$$\mathbf{F}\hat{\mathbf{a}} = \begin{pmatrix} p(1) \\ p(\xi) \\ \vdots \\ p(\xi^{2n-1}) \end{pmatrix},$$

and hence

$$\mathbf{F}(\mathbf{a} \odot \mathbf{b}) = \mathbf{D}\mathbf{F}\hat{\mathbf{b}} = (\mathbf{F}\hat{\mathbf{a}}) \times (\mathbf{F}\hat{\mathbf{b}}).$$

5.8.15. (a) $\mathbf{P}_n\mathbf{x}$ performs an even–odd permutation to all components of $\mathbf{x}$. The matrix

$$(\mathbf{I}_2 \otimes \mathbf{P}_{n/2}) = \begin{pmatrix} \mathbf{P}_{n/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{n/2} \end{pmatrix} \mathbf{x}$$

performs an even–odd permutation to the top half of $\mathbf{x}$ and then does the same to the bottom half of $\mathbf{x}$. The matrix

$$(\mathbf{I}_4 \otimes \mathbf{P}_{n/4}) = \begin{pmatrix} \mathbf{P}_{n/4} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{n/4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{n/4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_{n/4} \end{pmatrix} \mathbf{x}$$

performs an even–odd permutation to each individual quarter of $\mathbf{x}$. As this pattern is continued, the product

$$\mathbf{R}_n = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_{2^1})(\mathbf{I}_{2^{r-2}} \otimes \mathbf{P}_{2^2}) \cdots (\mathbf{I}_{2^1} \otimes \mathbf{P}_{2^{r-1}})(\mathbf{I}_{2^0} \otimes \mathbf{P}_{2^r})\mathbf{x}$$

produces the bit-reversing permutation. For example, when $n = 8$,

$$\mathbf{R}_8\mathbf{x} = (\mathbf{I}_4 \otimes \mathbf{P}_2)(\mathbf{I}_2 \otimes \mathbf{P}_4)(\mathbf{I}_1 \otimes \mathbf{P}_8)\mathbf{x}$$

$$= \begin{pmatrix} \mathbf{P}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_4 \end{pmatrix} \mathbf{P}_8 \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{P}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} \mathbf{P}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_4 \end{pmatrix} \begin{pmatrix} x_0 \\ x_2 \\ x_4 \\ x_6 \\ \hline x_1 \\ x_3 \\ x_5 \\ x_7 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{P}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{P}_2 \end{pmatrix} \begin{pmatrix} x_0 \\ x_4 \\ x_2 \\ x_6 \\ \hline x_1 \\ x_5 \\ x_3 \\ x_7 \end{pmatrix} = \begin{pmatrix} x_0 \\ x_4 \\ x_2 \\ x_6 \\ x_1 \\ x_5 \\ x_3 \\ x_7 \end{pmatrix} \quad \text{because} \quad \mathbf{P}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(b) To prove that $\mathbf{I}_{2^{r-k}} \otimes \mathbf{F}_{2^k} = \mathbf{L}_{2^k}\mathbf{R}_{2^k}$ using induction, note first that for $k = 1$ we have

$$\mathbf{L}_2 = (\mathbf{I}_{2^{r-1}} \otimes \mathbf{B}_2)1 = \mathbf{I}_{2^{r-1}} \otimes \mathbf{F}_2 \quad \text{and} \quad \mathbf{R}_2 = \mathbf{I}_n(\mathbf{I}_{2^{r-1}} \otimes \mathbf{P}_2) = \mathbf{I}_n\mathbf{I}_n = \mathbf{I}_n,$$

so $\mathbf{L}_2\mathbf{R}_2 = \mathbf{I}_{2^{r-1}} \otimes \mathbf{F}_2$. Now assume that the result holds for $k = j$—i.e., assume

$$\mathbf{I}_{2^{r-j}} \otimes \mathbf{F}_{2^j} = \mathbf{L}_{2^j}\mathbf{R}_{2^j}.$$

Prove that the result is true for $k = j + 1$—i.e., prove

$$\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{F}_{2^{j+1}} = \mathbf{L}_{2^{j+1}}\mathbf{R}_{2^{j+1}}.$$

Use the fact that $\mathbf{F}_{2^{j+1}} = \mathbf{B}_{2^{j+1}}(\mathbf{I}_2 \otimes \mathbf{F}_j)\mathbf{P}_{2^{j+1}}$ along with the two basic prop-

erties of the tensor product given in the introduction of this exercise to write

$$\begin{aligned}
\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{F}_{2^{j+1}} &= \mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{B}_{2^{j+1}}(\mathbf{I}_2 \otimes \mathbf{F}_{2^j})\mathbf{P}_{2^{j+1}} \\
&= \Big(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{B}_{2^{j+1}}(\mathbf{I}_2 \otimes \mathbf{F}_{2^j})\Big)\Big(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{P}_{2^{j+1}}\Big) \\
&= (\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{B}_{2^{j+1}})(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{I}_2 \otimes \mathbf{F}_{2^j})(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{P}_{2^{j+1}}) \\
&= (\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{B}_{2^{j+1}})(\mathbf{I}_{2^{r-j}} \otimes \mathbf{F}_{2^j})(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{P}_{2^{j+1}}) \\
&= (\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{B}_{2^{j+1}})\mathbf{L}_{2^j}\mathbf{R}_{2^j}(\mathbf{I}_{2^{r-(j+1)}} \otimes \mathbf{P}_{2^{j+1}}) \\
&= \mathbf{L}_{2^{j+1}}\mathbf{R}_{2^{j+1}}.
\end{aligned}$$

Therefore, $\mathbf{I}_{2^{r-k}} \otimes \mathbf{F}_{2^k} = \mathbf{L}_{2^k}\mathbf{R}_{2^k}$ for $k = 1, 2, \ldots, r$, and when $k = r$ we have that $\mathbf{F}_n = \mathbf{L}_n\mathbf{R}_n$.

5.8.16. According to Exercise 5.8.10,

$$\mathbf{F}_n\mathbf{a} = \mathbf{b}, \quad \text{where} \quad \mathbf{a} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{n-1} \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} p(1) \\ p(\xi) \\ p(\xi^2) \\ \vdots \\ p(\xi^{n-1}) \end{pmatrix}.$$

By making use of the fact that $(1/\sqrt{n})\mathbf{F}_n$ is unitary we can write

$$\sum_{k=0}^{n-1} \left|p(\xi^k)\right|^2 = \mathbf{b}^*\mathbf{b} = (\mathbf{F}_n\mathbf{a})^*(\mathbf{F}_n\mathbf{a}) = \mathbf{a}^*\mathbf{F}_n^*\mathbf{F}_n\mathbf{a} = \mathbf{a}^*(n\mathbf{I})\mathbf{a} = n\sum_{k=0}^{n-1} |\alpha_k|^2.$$

5.8.17. Let $\mathbf{y} = (2/n)\mathbf{F}\mathbf{x}$, and use the result in (5.8.7) to write

$$\begin{aligned}
\|\mathbf{y}\|^2 &= \left\|\sum_k \Big((\alpha_k - \mathrm{i}\beta_k)\mathbf{e}_{f_k} + (\alpha_k + \mathrm{i}\beta_k)\mathbf{e}_{n-f_k}\Big)\right\| \\
&= \sum_k \Big(|\alpha_k - \mathrm{i}\beta_k|^2 + |\alpha_k + \mathrm{i}\beta_k|^2\Big) \\
&= 2\sum_k \Big(\alpha_k^2 + \beta_k^2\Big).
\end{aligned}$$

But because $\mathbf{F}^*\mathbf{F} = n\mathbf{I}$, it follows that

$$\|\mathbf{y}\|^2 = \left\|\frac{2}{n}\mathbf{F}\mathbf{x}\right\|^2 = \frac{4}{n^2}\mathbf{x}^*\mathbf{F}^*\mathbf{F}\mathbf{x} = \frac{4}{n}\|\mathbf{x}\|^2,$$

so combining these two statements produces the desired conclusion.

5.8.18. We know from (5.8.11) that if $p(x) = \sum_{k=0}^{n-1} \alpha_k x^k$, then

$$p^2(x) = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{a}]_k x^k.$$

The last component of $\mathbf{a} \odot \mathbf{a}$ is zero, so we can write

$$\mathbf{c}^T(\mathbf{a} \odot \mathbf{a}) = \sum_{k=0}^{2n-2} [\mathbf{a} \odot \mathbf{a}]_k \eta^k = p^2(\eta) = \left(\sum_{k=0}^{n-1} \alpha_k \eta^k \right)^2 = \left(\mathbf{c}^T \hat{\mathbf{a}}\right)^2.$$

5.8.19. Start with $\mathbf{X} \longleftarrow rev(\mathbf{x}) = (x_0\, x_4\, x_2\, x_6\, x_1\, x_5\, x_3\, x_7)$.

For $j = 0$:

$$\mathbf{D} \longleftarrow (1)$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 & x_2 & x_1 & x_3 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_4 & x_6 & x_5 & x_7 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} x_0 + x_4 & x_2 + x_6 & x_1 + x_5 & x_3 + x_7 \\ x_0 - x_4 & x_2 - x_6 & x_1 - x_5 & x_3 - x_7 \end{pmatrix}_{2\times 8}$$

For $j = 1$:

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ \mathrm{e}^{-\pi \mathrm{i}/2} \end{pmatrix} = \begin{pmatrix} 1 \\ \xi^2 \end{pmatrix}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 + x_4 & x_1 + x_5 \\ x_0 - x_4 & x_1 - x_5 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_2 + x_6 & x_3 + x_7 \\ x_2 - x_6 & x_3 - x_7 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix}$$

$$= \left(\begin{array}{c|c} x_0 + x_4 + \quad x_2 + \quad x_6 & x_1 + x_5 + \quad x_3 + \quad x_7 \\ x_0 - x_4 + \xi^2 x_2 - \xi^2 x_6 & x_1 - x_5 + \xi^2 x_3 - \xi^2 x_7 \\ x_0 + x_4 - \quad x_2 - \quad x_6 & x_1 + x_5 - \quad x_3 - \quad x_7 \\ x_0 - x_4 - \xi^2 x_2 + \xi^2 x_6 & x_1 - x_5 - \xi^2 x_3 + \xi^2 x_7 \end{array}\right)_{4\times 2}$$

For $j = 2$:

$$\mathbf{D} \longleftarrow \begin{pmatrix} 1 \\ \mathrm{e}^{-\pi \mathrm{i}/4} \\ \mathrm{e}^{-2\pi \mathrm{i}/4} \\ \mathrm{e}^{-3\pi \mathrm{i}/4} \end{pmatrix} = \begin{pmatrix} 1 \\ \xi \\ \xi^2 \\ \xi^3 \end{pmatrix}$$

$$\mathbf{X}^{(0)} \longleftarrow \begin{pmatrix} x_0 + x_4 + x_2 + x_6 \\ x_0 - x_4 + \xi^2 x_2 - \xi^2 x_6 \\ x_0 + x_4 - x_2 - x_6 \\ x_0 - x_4 - \xi^2 x_2 + \xi^2 x_6 \end{pmatrix}$$

$$\mathbf{X}^{(1)} \longleftarrow \begin{pmatrix} x_1 + x_5 + x_3 + x_7 \\ x_1 - x_5 + \xi^2 x_3 - \xi^2 x_7 \\ x_1 + x_5 - x_3 - x_7 \\ x_1 - x_5 - \xi^2 x_3 + \xi^2 x_7 \end{pmatrix}$$

$$\mathbf{X} \longleftarrow \begin{pmatrix} \mathbf{X}^{(0)} + \mathbf{D} \times \mathbf{X}^{(1)} \\ \mathbf{X}^{(0)} - \mathbf{D} \times \mathbf{X}^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} x_0 + x_4 + x_2 + x_6 + x_1 + x_5 + x_3 + x_7 \\ x_0 - x_4 + \xi^2 x_2 - \xi^2 x_6 + \xi\, x_1 - \xi x_5 + \xi^3 x_3 - \xi^3 x_7 \\ x_0 + x_4 - x_2 - x_6 + \xi^2 x_1 + \xi^2 x_5 - \xi^2 x_3 - \xi^2 x_7 \\ x_0 - x_4 - \xi^2 x_2 + \xi^2 x_6 + \xi^3 x_1 - \xi^3 x_5 - \xi^5 x_3 + \xi^5 x_7 \\ x_0 + x_4 + x_2 + x_6 - x_1 - x_5 - x_3 - x_7 \\ x_0 - x_4 + \xi^2 x_2 - \xi^2 x_6 - \xi\, x_1 + \xi\, x_5 - \xi^3 x_3 + \xi^3 x_7 \\ x_0 + x_4 - x_2 - x_6 - \xi^2 x_1 - \xi^2 x_5 + \xi^2 x_3 + \xi^2 x_7 \\ x_0 - x_4 - \xi^2 x_2 + \xi^2 x_6 - \xi^3 x_1 + \xi^3 x_5 + \xi^5 x_3 - \xi^5 x_7 \end{pmatrix}_{8\times 1}$$

To verify that this is the same as $\mathbf{F}_8\mathbf{x}_8$, use the fact that $\xi = -\xi^5$, $\xi^2 = -\xi^6$, $\xi^3 = -\xi^7$, and $\xi^4 = -1$.

Solutions for exercises in section 5. 9

5.9.1. (a) The fact that

$$rank\,(\mathbf{B}) = rank\big[\mathbf{X}\,|\,\mathbf{Y}\big] = rank\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} = 3$$

implies $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is a basis for $\Re^3$, so (5.9.4) guarantees that $\mathcal{X}$ and $\mathcal{Y}$ are complementary.

(b) According to (5.9.12), the projector onto $\mathcal{X}$ along $\mathcal{Y}$ is

$$\mathbf{P} = \big[\mathbf{X}\,|\,\mathbf{0}\big]\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}^{-1}$$

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

and (5.9.9) insures that the complementary projector onto $\mathcal{Y}$ along $\mathcal{X}$ is

$$\mathbf{Q} = \mathbf{I} - \mathbf{P} = \begin{pmatrix} 0 & -1 & 1 \\ 0 & -2 & 2 \\ 0 & -3 & 3 \end{pmatrix}.$$

(c) $\mathbf{Qv} = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$

(d) Direct multiplication shows $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{Q}^2 = \mathbf{Q}$.

(e) To verify that $R(\mathbf{P}) = \mathcal{X} = N(\mathbf{Q})$, you can use the technique of Example 4.2.2 to show that the basic columns of $\mathbf{P}$ (or the columns in a basis for $N(\mathbf{Q})$) span the same space generated by $\mathcal{B}_{\mathcal{X}}$. To verify that $N(\mathbf{P}) = \mathcal{Y}$, note that $\mathbf{P}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ together with the fact that $\dim N(\mathbf{P}) = 3 - rank(\mathbf{P}) = 1$.

5.9.2. There are many ways to do this. One way is to write down any basis for $\Re^5$—say $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$—and set

$$\mathcal{X} = span\{\mathbf{x}_1, \mathbf{x}_2\} \quad \text{and} \quad \mathcal{Y} = span\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}.$$

Property (5.9.4) guarantees that $\mathcal{X}$ and $\mathcal{Y}$ are complementary.

5.9.3. Let $\mathcal{X} = \{(\alpha, \alpha) \,|\, \alpha \in \Re\}$ and $\mathcal{Y} = \Re^2$ so that $\Re^2 = \mathcal{X} + \mathcal{Y}$, but $\mathcal{X} \cap \mathcal{Y} \neq \mathbf{0}$. For each vector in $\Re^2$ we can write

$$(x, y) = (x, x) + (0, y - x) \quad \text{and} \quad (x, y) = (y, y) + (x - y, 0).$$

5.9.4. Exercise 3.2.6 says that each $\mathbf{A} \in \Re^{n\times n}$ can be uniquely written as the sum of a symmetric matrix and a skew-symmetric matrix according to the formula

$$\mathbf{A} = \frac{\mathbf{A} + \mathbf{A}^T}{2} + \frac{\mathbf{A} - \mathbf{A}^T}{2},$$

so (5.9.3) guarantees that $\Re^{n\times n} = \mathcal{S} \oplus \mathcal{K}$. By definition, the projection of $\mathbf{A}$ onto $\mathcal{S}$ along $\mathcal{K}$ is the $\mathcal{S}$-component of $\mathbf{A}$—namely $(\mathbf{A} + \mathbf{A}^T)/2$. For the given matrix, this is

$$\frac{\mathbf{A} + \mathbf{A}^T}{2} = \begin{pmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{pmatrix}.$$

5.9.5. (a) Assume that $\mathcal{X} \cap \mathcal{Y} = \mathbf{0}$. To prove $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is linearly independent, write

$$\begin{aligned}
\sum_{i=1}^{m} \alpha_i \mathbf{x}_i + \sum_{j=1}^{n} \beta_j \mathbf{y}_j = \mathbf{0} \quad &\Longrightarrow \quad \sum_{i=1}^{m} \alpha_i \mathbf{x}_i = -\sum_{j=1}^{n} \beta_j \mathbf{y}_j \\
&\Longrightarrow \quad \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \in \mathcal{X} \cap \mathcal{Y} = \mathbf{0} \\
&\Longrightarrow \quad \sum_{i=1}^{m} \alpha_i \mathbf{x}_i = \mathbf{0} \quad \text{and} \quad \sum_{j=1}^{n} \beta_j \mathbf{y}_j = \mathbf{0} \\
&\Longrightarrow \quad \alpha_1 = \cdots = \alpha_m = \beta_1 = \cdots = \beta_n = 0 \\
&\qquad \text{(because } \mathcal{B}_{\mathcal{X}} \text{ and } \mathcal{B}_{\mathcal{Y}} \text{ are both independent).}
\end{aligned}$$

Conversely, if $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is linearly independent, then

$$\begin{aligned}
\mathbf{v} \in \mathcal{X} \cap \mathcal{Y} \quad &\Longrightarrow \quad \mathbf{v} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \quad \text{and} \quad \mathbf{v} = \sum_{j=1}^{n} \beta_j \mathbf{y}_j \\
&\Longrightarrow \quad \sum_{i=1}^{m} \alpha_i \mathbf{x}_i - \sum_{j=1}^{n} \beta_j \mathbf{y}_j = \mathbf{0} \\
&\Longrightarrow \quad \alpha_1 = \cdots = \alpha_m = \beta_1 = \cdots = \beta_n = 0 \\
&\qquad\quad \text{(because } \mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}} \text{ is independent)} \\
&\Longrightarrow \quad \mathbf{v} = \mathbf{0}.
\end{aligned}$$

(b) No. Take $\mathcal{X}$ to be the xy-plane and $\mathcal{Y}$ to be the yz-plane in $\Re^3$ with $\mathcal{B}_{\mathcal{X}} = \{\mathbf{e}_1, \mathbf{e}_2\}$ and $\mathcal{B}_{\mathcal{Y}} = \{\mathbf{e}_2, \mathbf{e}_3\}$. We have $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, but $\mathcal{X} \cap \mathcal{Y} \neq \mathbf{0}$.

(c) No, the fact that $\mathcal{B}_{\mathcal{X}} \cup \mathcal{B}_{\mathcal{Y}}$ is linearly independent is no guarantee that $\mathcal{X} + \mathcal{Y}$ is the entire space—e.g., consider two distinct lines in $\Re^3$.

5.9.6. If $\mathbf{x}$ is a fixed point for $\mathbf{P}$, then $\mathbf{Px} = \mathbf{x}$ implies $\mathbf{x} \in R(\mathbf{P})$. Conversely, if $\mathbf{x} \in R(\mathbf{P})$, then $\mathbf{x} = \mathbf{Py}$ for some $\mathbf{y} \in \mathcal{V} \implies \mathbf{Px} = \mathbf{P}^2\mathbf{y} = \mathbf{Py} = \mathbf{x}$.

5.9.7. Use (5.9.10) (which you just validated in Exercise 5.9.6) in conjunction with the definition of a projector onto $\mathcal{X}$ to realize that

$$\mathbf{x} \in \mathcal{X} \Longleftrightarrow \mathbf{Px} = \mathbf{x} \Longleftrightarrow \mathbf{x} \in R(\mathbf{P}),$$

and

$$\mathbf{x} \in R(\mathbf{P}) \Longleftrightarrow \mathbf{Px} = \mathbf{x} \Longleftrightarrow (\mathbf{I} - \mathbf{P})\mathbf{x} = \mathbf{0} \Longleftrightarrow \mathbf{x} \in N(\mathbf{I} - \mathbf{P}).$$

The statements concerning the complementary projector $\mathbf{I} - \mathbf{P}$ are proven in a similar manner.

5.9.8. If θ is the angle between $R(\mathbf{P})$ and $N(\mathbf{P})$, it follows from (5.9.18) that $\|\mathbf{P}\|_2 = (1/\sin\theta) \geq 1$. Furthermore, $\|\mathbf{P}\|_2 = 1$ if and only if $\sin\theta = 1$ (i.e., $\theta = \pi/2$), which is equivalent to saying $R(\mathbf{P}) \perp N(\mathbf{P})$.

5.9.9. Let θ be the angle between $R(\mathbf{P})$ and $N(\mathbf{P})$. We know from (5.9.11) that $R(\mathbf{I} - \mathbf{P}) = N(\mathbf{P})$ and $N(\mathbf{I} - \mathbf{P}) = R(\mathbf{P})$, so θ is also the angle between $R(\mathbf{I} - \mathbf{P})$ and $N(\mathbf{I} - \mathbf{P})$. Consequently, (5.9.18) says that

$$\|\mathbf{I} - \mathbf{P}\|_2 = \frac{1}{\sin\theta} = \|\mathbf{P}\|_2 .$$

5.9.10. The trick is to observe that $\mathbf{P} = \mathbf{uv}^T$ is a projector because $\mathbf{v}^T\mathbf{u} = 1$ implies $\mathbf{P}^2 = \mathbf{uv}^T\mathbf{uv}^T = \mathbf{uv}^T = \mathbf{P}$, so the result of Exercise 5.9.9 insures that

$$\left\|\mathbf{I} - \mathbf{uv}^T\right\|_2 = \left\|\mathbf{uv}^T\right\|_2 .$$

To prove that $\left\|\mathbf{u}\mathbf{v}^T\right\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$, start with the definition of an induced matrix given in (5.2.4) on p. 280, and write $\left\|\mathbf{u}\mathbf{v}^T\right\|_2 = \max_{\|\mathbf{x}\|_2=1} \left\|\mathbf{u}\mathbf{v}^T\mathbf{x}\right\|_2$. If the maximum occurs at $\mathbf{x} = \mathbf{x}_0$ with $\|\mathbf{x}_0\|_2 = 1$, then

$$\begin{aligned}\left\|\mathbf{u}\mathbf{v}^T\right\|_2 &= \left\|\mathbf{u}\mathbf{v}^T\mathbf{x}_0\right\|_2 = \|\mathbf{u}\|_2 \left|\mathbf{v}^T\mathbf{x}_0\right| \\ &\le \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \|\mathbf{x}_0\|_2 \text{ by CBS inequality} \\ &= \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 .\end{aligned}$$

But we can also write

$$\begin{aligned}\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 &= \|\mathbf{u}\|_2 \frac{\|\mathbf{v}\|_2^2}{\|\mathbf{v}\|_2} = \|\mathbf{u}\|_2 \frac{(\mathbf{v}^T\mathbf{v})}{\|\mathbf{v}\|_2} = \frac{\left\|\mathbf{u}\mathbf{v}^T\mathbf{v}\right\|_2}{\|\mathbf{v}\|_2} \\ &= \left\|\mathbf{u}\mathbf{v}^T\left(\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right)\right\|_2 \le \max_{\|\mathbf{x}\|_2=1} \left\|\mathbf{u}\mathbf{v}^T\mathbf{x}\right\|_2 \\ &= \left\|\mathbf{u}\mathbf{v}^T\right\|_2 ,\end{aligned}$$

so $\left\|\mathbf{u}\mathbf{v}^T\right\|_2 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$. Finally, if $\mathbf{P} = \mathbf{u}\mathbf{v}^T$, use Example 3.6.5 to write

$$\|\mathbf{P}\|_F^2 = trace\left(\mathbf{P}^T\mathbf{P}\right) = trace(\mathbf{v}\mathbf{u}^T\mathbf{u}\mathbf{v}^T) = trace(\mathbf{u}^T\mathbf{u}\mathbf{v}^T\mathbf{v}) = \|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 .$$

5.9.11. $\mathbf{p} = \mathbf{P}\mathbf{v} = [\mathbf{X} \,|\, \mathbf{0}][\mathbf{X} \,|\, \mathbf{Y}]^{-1}\mathbf{v} = [\mathbf{X} \,|\, \mathbf{0}]\mathbf{z} = \mathbf{X}\mathbf{z}_1$

5.9.12. (a) Use (5.9.10) to conclude that

$$\begin{aligned}R(\mathbf{P}) = R(\mathbf{Q}) &\implies \mathbf{P}\mathbf{Q}_{*j} = \mathbf{Q}_{*j} \quad \text{and} \quad \mathbf{Q}\mathbf{P}_{*j} = \mathbf{P}_{*j} \quad \forall\ j \\ &\implies \mathbf{P}\mathbf{Q} = \mathbf{Q} \quad \text{and} \quad \mathbf{Q}\mathbf{P} = \mathbf{P}.\end{aligned}$$

Conversely, use Exercise 4.2.12 to write

$$\left\{\begin{array}{lll}\mathbf{P}\mathbf{Q} = \mathbf{Q} & \implies & R(\mathbf{Q}) \subseteq R(\mathbf{P}) \\ \mathbf{Q}\mathbf{P} = \mathbf{P} & \implies & R(\mathbf{P}) \subseteq R(\mathbf{Q})\end{array}\right\} \implies R(\mathbf{P}) = R(\mathbf{Q}).$$

(b) Use $N(\mathbf{P}) = N(\mathbf{Q}) \iff R(\mathbf{I} - \mathbf{P}) = R(\mathbf{I} - \mathbf{Q})$ together with part (a).

(c) From part (a), $\mathbf{E}_i\mathbf{E}_j = \mathbf{E}_j$ so that

$$\begin{aligned}\Big(\sum_j \alpha_j\mathbf{E}_j\Big)^2 &= \sum_i\sum_j \alpha_i\alpha_j\mathbf{E}_i\mathbf{E}_j = \sum_i\sum_j \alpha_i\alpha_j\mathbf{E}_j \\ &= \Big(\sum_i \alpha_i\Big)\Big(\sum_j \alpha_j\mathbf{E}_j\Big) = \sum_j \alpha_j\mathbf{E}_j.\end{aligned}$$

5.9.13. According to (5.9.12), the projector onto $\mathcal{X}$ along $\mathcal{Y}$ is $\mathbf{P} = \mathbf{B}\begin{pmatrix}\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{B}^{-1}$, where $\mathbf{B} = [\mathbf{X} \,|\, \mathbf{Y}]$ in which the columns of $\mathbf{X}$ and $\mathbf{Y}$ form bases for $\mathcal{X}$

and $\mathcal{Y}$, respectively. Since multiplication by nonsingular matrices does not alter the rank, it follows that $rank\,(\mathbf{P}) = rank\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = r$. Using the result of Example 3.6.5 produces

$$\begin{aligned} trace\,(\mathbf{P}) &= trace\left[\mathbf{B}\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{B}^{-1}\right] = trace\left[\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{B}^{-1}\mathbf{B}\right]\\ &= trace\begin{pmatrix}\mathbf{I}_r & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = r = rank\,(\mathbf{P}). \end{aligned}$$

5.9.14. **(i)** $\Longrightarrow$ **(ii)** : If $\mathbf{v} = \mathbf{x}_1 + \cdots + \mathbf{x}_k$ and $\mathbf{v} = \mathbf{y}_1 + \cdots + \mathbf{y}_k$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathcal{X}_i$, then

$$\begin{aligned} \sum_{i=1}^{k}(\mathbf{x}_i - \mathbf{y}_i) = \mathbf{0} \quad &\Longrightarrow \quad (\mathbf{x}_k - \mathbf{y}_k) = -\sum_{i=1}^{k-1}(\mathbf{x}_i - \mathbf{y}_i)\\ &\Longrightarrow \quad (\mathbf{x}_k - \mathbf{y}_k) \in \mathcal{X}_k \cap (\mathcal{X}_1 + \cdots + \mathcal{X}_{k-1}) = \mathbf{0}\\ &\Longrightarrow \quad \mathbf{x}_k = \mathbf{y}_k \quad \text{and} \quad \sum_{i=1}^{k-1}(\mathbf{x}_i - \mathbf{y}_i) = \mathbf{0}. \end{aligned}$$

Now repeat the argument—to be formal, use induction.

(ii) $\Longrightarrow$ **(iii)** : The proof is essentially the same argument as that used to establish (5.9.3) $\Longrightarrow$ (5.9.4).

(iii) $\Longrightarrow$ **(i)** : $\mathcal{B}$ always spans $\mathcal{X}_1 + \mathcal{X}_2 + \cdots + \mathcal{X}_k$, and since the hypothesis is that $\mathcal{B}$ is a basis for $\mathcal{V}$, it follows that $\mathcal{B}$ is a basis for both $\mathcal{V}$ and $\mathcal{X}_1 + \cdots + \mathcal{X}_k$. Consequently $\mathcal{V} = \mathcal{X}_1 + \mathcal{X}_2 + \cdots + \mathcal{X}_k$. Furthermore, the set $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{k-1}$ is linearly independent (each subset of an independent set is independent), and it spans $\mathcal{V}_{k-1} = \mathcal{X}_1 + \cdots + \mathcal{X}_{k-1}$, so $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{k-1}$ must be a basis for $\mathcal{V}_{k-1}$. Now, since $(\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_{k-1}) \cup \mathcal{B}_k$ is a basis for $\mathcal{V} = (\mathcal{X}_1 + \cdots + \mathcal{X}_{k-1}) + \mathcal{X}_k$, it follows from (5.9.2)–(5.9.4) that $\mathcal{V} = (\mathcal{X}_1 + \cdots + \mathcal{X}_{k-1}) \oplus \mathcal{X}_k$, so $\mathcal{X}_k \cap (\mathcal{X}_1 + \cdots + \mathcal{X}_{k-1}) = \mathbf{0}$. The same argument can now be repeated on $\mathcal{V}_{k-1}$—to be formal, use induction.

5.9.15. We know from (5.9.12) that $\mathbf{P} = \mathbf{Q}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}$ and $\mathbf{I} - \mathbf{P} = \mathbf{Q}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\mathbf{Q}^{-1}$, so

$$\begin{aligned} \mathbf{PAP} &= \mathbf{Q}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}\mathbf{Q}\begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix}\mathbf{Q}^{-1}\mathbf{Q}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}\\ &= \mathbf{Q}\begin{pmatrix}\mathbf{A}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}. \end{aligned}$$

The other three statements are derived in an analogous fashion.

5.9.16. According to (5.9.12), the projector onto $\mathcal{X}$ along $\mathcal{Y}$ is $\mathbf{P} = \big[\mathbf{X}\,|\,\mathbf{0}\big]\big[\mathbf{X}\,|\,\mathbf{Y}\big]^{-1}$, where the columns of $\mathbf{X}$ and $\mathbf{Y}$ are bases for $\mathcal{X}$ and $\mathcal{Y}$, respectively. If

$$\left[\mathbf{X}_{n\times r}\,|\,\mathbf{Y}\right]^{-1} = \begin{pmatrix} \mathbf{A}_{r\times n} \\ \mathbf{C} \end{pmatrix}, \text{ then}$$

$$\mathbf{P} = \left[\mathbf{X}_{n\times r}\,|\,\mathbf{0}\right]\begin{pmatrix} \mathbf{A}_{r\times n} \\ \mathbf{C} \end{pmatrix} = \mathbf{X}_{n\times r}\mathbf{A}_{r\times n}.$$

The nonsingularity of $\left[\mathbf{X}\,|\,\mathbf{Y}\right]$ and $\begin{pmatrix} \mathbf{A} \\ \mathbf{C} \end{pmatrix}$ insures that $\mathbf{X}$ has full column rank and $\mathbf{A}$ has full row rank. The fact that $\mathbf{AX} = \mathbf{I}_r$ is a consequence of

$$\begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \left[\mathbf{X}\,|\,\mathbf{Y}\right]^{-1}\left[\mathbf{X}\,|\,\mathbf{Y}\right] = \begin{pmatrix} \mathbf{A}_{r\times n} \\ \mathbf{C} \end{pmatrix}\left[\mathbf{X}_{n\times r}\,|\,\mathbf{Y}\right] = \begin{pmatrix} \mathbf{AX} & \mathbf{AY} \\ \mathbf{CX} & \mathbf{CY} \end{pmatrix}.$$

5.9.17. (a) Use the fact that a linear operator $\mathbf{P}$ is a projector if and only if $\mathbf{P}$ is idempotent. If $\mathbf{EF} = \mathbf{FE} = \mathbf{0}$, then $(\mathbf{E}+\mathbf{F})^2 = \mathbf{E}+\mathbf{F}$. Conversely, if $\mathbf{E}+\mathbf{F}$ is a projector, then

$$\begin{aligned}
(\mathbf{E}+\mathbf{F})^2 = \mathbf{E}+\mathbf{F} \quad &\Longrightarrow \quad \mathbf{EF}+\mathbf{FE} = \mathbf{0} \\
&\Longrightarrow \quad \mathbf{E}(\mathbf{EF}+\mathbf{FE}) = \mathbf{0} \quad \text{and} \quad (\mathbf{EF}+\mathbf{FE})\mathbf{E} = \mathbf{0} \\
&\Longrightarrow \quad \mathbf{EF} = \mathbf{FE} \\
&\Longrightarrow \quad \mathbf{EF} = \mathbf{0} = \mathbf{FE} \quad (\text{because } \mathbf{EF}+\mathbf{FE} = \mathbf{0}).
\end{aligned}$$

Thus $\mathbf{P} = \mathbf{E}+\mathbf{F}$ is a projector if and only if $\mathbf{EF} = \mathbf{FE} = \mathbf{0}$. Now prove that under this condition $R(\mathbf{P}) = \mathcal{X}_1 \oplus \mathcal{X}_2$. Start with the fact that $\mathbf{z} \in R(\mathbf{P})$ if and only if $\mathbf{Pz} = \mathbf{z}$, and write each such vector $\mathbf{z}$ as $\mathbf{z} = \mathbf{x}_1 + \mathbf{y}_1$ and $\mathbf{z} = \mathbf{x}_2 + \mathbf{y}_2$, where $\mathbf{x}_i \in \mathcal{X}_i$ and $\mathbf{y}_i \in \mathcal{Y}_i$ so that $\mathbf{Ex}_1 = \mathbf{x}_1$, $\mathbf{Ey}_1 = \mathbf{0}$, $\mathbf{Fx}_2 = \mathbf{x}_2$, and $\mathbf{Fy}_2 = \mathbf{0}$. To prove that $R(\mathbf{P}) = \mathcal{X}_1 + \mathcal{X}_2$, write

$$\begin{aligned}
\mathbf{z} \in R(\mathbf{P}) \quad &\Longrightarrow \quad \mathbf{Pz} = \mathbf{z} \quad \Longrightarrow \quad (\mathbf{E}+\mathbf{F})\mathbf{z} = \mathbf{z} \\
&\Longrightarrow \quad (\mathbf{E}+\mathbf{F})(\mathbf{x}_2+\mathbf{y}_2) = (\mathbf{x}_2+\mathbf{y}_2) \\
&\Longrightarrow \quad \mathbf{Ez} = \mathbf{y}_2 \quad \Longrightarrow \quad \mathbf{x}_1 = \mathbf{y}_2 \\
&\Longrightarrow \quad \mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2 \in \mathcal{X}_1 + \mathcal{X}_2 \quad \Longrightarrow \quad R(\mathbf{P}) \subseteq \mathcal{X}_1 + \mathcal{X}_2.
\end{aligned}$$

Conversely, $\mathcal{X}_1 + \mathcal{X}_2 \subseteq R(\mathbf{P})$ because

$$\begin{aligned}
\mathbf{z} \in \mathcal{X}_1 + \mathcal{X}_2 \quad &\Longrightarrow \quad \mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2, \quad \text{where} \quad \mathbf{x}_1 \in \mathcal{X}_1 \text{ and } \mathbf{x}_2 \in \mathcal{X}_2 \\
&\Longrightarrow \quad \mathbf{x}_1 = \mathbf{Ex}_1 \quad \text{and} \quad \mathbf{x}_2 = \mathbf{Fx}_2 \\
&\Longrightarrow \quad \mathbf{Fx}_1 = \mathbf{FEx}_1 = \mathbf{0} \text{ and } \mathbf{Ex}_2 = \mathbf{EFx}_2 = \mathbf{0} \\
&\Longrightarrow \quad \mathbf{Pz} = (\mathbf{E}+\mathbf{F})(\mathbf{x}_1+\mathbf{x}_2) = \mathbf{x}_1 + \mathbf{x}_2 = \mathbf{z} \\
&\Longrightarrow \quad \mathbf{z} \in R(\mathbf{P}).
\end{aligned}$$

The fact that $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint follows by writing

$$\mathbf{z} \in \mathcal{X}_1 \cap \mathcal{X}_2 \quad \Longrightarrow \quad \mathbf{Ez} = \mathbf{z} = \mathbf{Fz} \quad \Longrightarrow \quad \mathbf{z} = \mathbf{EFz} = \mathbf{0},$$

and thus $R(\mathbf{P}) = \mathcal{X}_1 \oplus \mathcal{X}_2$ is established. To prove that $N(\mathbf{P}) = \mathcal{Y}_1 \cap \mathcal{Y}_2$, write

$$\begin{aligned}
\mathbf{Pz} = 0 \quad &\Longrightarrow \quad (\mathbf{E}+\mathbf{F})\mathbf{z} = 0 \quad \Longrightarrow \quad \mathbf{Ez} = -\mathbf{Fz} \\
&\Longrightarrow \quad \mathbf{Ez} = -\mathbf{EFz} \quad \text{and} \quad \mathbf{FEz} = -\mathbf{Fz} \\
&\Longrightarrow \quad \mathbf{Ez} = \mathbf{0} \quad \text{and} \quad \mathbf{0} = \mathbf{Fz} \quad \Longrightarrow \quad \mathbf{z} \in \mathcal{Y}_1 \cap \mathcal{Y}_2.
\end{aligned}$$

5.9.18. Use the hint together with the result of Exercise 5.9.17 to write

$$\begin{aligned}
\mathbf{E}-\mathbf{F} \text{ is a projector} &\Longleftrightarrow \mathbf{I}-(\mathbf{E}-\mathbf{F}) \text{ is a projector} \\
&\Longleftrightarrow (\mathbf{I}-\mathbf{E})+\mathbf{F} \text{ is a projector} \\
&\Longleftrightarrow (\mathbf{I}-\mathbf{E})\mathbf{F} = \mathbf{0} = \mathbf{F}(\mathbf{I}-\mathbf{E}) \\
&\Longleftrightarrow \mathbf{EF} = \mathbf{F} = \mathbf{FE}.
\end{aligned}$$

Under this condition, Exercise 5.9.17 says that

$$R(\mathbf{I}-\mathbf{E}+\mathbf{F}) = R(\mathbf{I}-\mathbf{E}) \oplus R(\mathbf{F}) \text{ and } N(\mathbf{I}-\mathbf{E}+\mathbf{F}) = N(\mathbf{I}-\mathbf{E}) \cap N(\mathbf{F}),$$

so (5.9.11) guarantees

$$\begin{aligned}
R(\mathbf{E}-\mathbf{F}) &= N(\mathbf{I}-\mathbf{E}+\mathbf{F}) = N(\mathbf{I}-\mathbf{E}) \cap N(\mathbf{F}) = R(\mathbf{E}) \cap N(\mathbf{F}) = \mathcal{X}_1 \cap \mathcal{Y}_2 \\
N(\mathbf{E}-\mathbf{F}) &= R(\mathbf{I}-\mathbf{E}+\mathbf{F}) = R(\mathbf{I}-\mathbf{E}) \oplus R(\mathbf{F}) = N(\mathbf{E}) \oplus R(\mathbf{F}) = \mathcal{Y}_1 \oplus \mathcal{X}_2.
\end{aligned}$$

5.9.19. If $\mathbf{EF} = \mathbf{P} = \mathbf{FE}$, then $\mathbf{P}$ is idempotent, and hence $\mathbf{P}$ is a projector. To prove that $R(\mathbf{P}) = \mathcal{X}_1 \cap \mathcal{X}_2$, write

$$\begin{aligned}
\mathbf{z} \in R(\mathbf{P}) \quad &\Longrightarrow \quad \mathbf{Pz} = \mathbf{z} \\
&\Longrightarrow \quad \mathbf{E}(\mathbf{Fz}) = \mathbf{z} \quad \text{and} \quad \mathbf{F}(\mathbf{Ez}) = \mathbf{z} \\
&\Longrightarrow \quad \mathbf{z} \in R(\mathbf{E}) \cap R(\mathbf{F}) = \mathcal{X}_1 \cap \mathcal{X}_2 \\
&\Longrightarrow \quad R(\mathbf{P}) \subseteq \mathcal{X}_1 \cap \mathcal{X}_2.
\end{aligned}$$

Conversely,

$$\mathbf{z} \in \mathcal{X}_1 \cap \mathcal{X}_2 \quad \Longrightarrow \quad \mathbf{Ez} = \mathbf{z} = \mathbf{Fz} \quad \Longrightarrow \quad \mathbf{Pz} = \mathbf{z} \quad \Longrightarrow \quad \mathcal{X}_1 \cap \mathcal{X}_2 \subseteq R(\mathbf{P}),$$

and hence $R(\mathbf{P}) = \mathcal{X}_1 \cap \mathcal{X}_2$. To prove that $N(\mathbf{P}) = \mathcal{Y}_1 + \mathcal{Y}_2$, first notice that

$$\mathbf{z} \in N(\mathbf{P}) \quad \Longrightarrow \quad \mathbf{FEz} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{Ez} \in N(\mathbf{F}).$$

This together with the fact that $(\mathbf{I}-\mathbf{E})\mathbf{z} \in N(\mathbf{E})$ allows us to conclude that

$$\mathbf{z} = (\mathbf{I}-\mathbf{E})\mathbf{z} + \mathbf{Ez} \in N(\mathbf{E}) + N(\mathbf{F}) = \mathcal{Y}_1 + \mathcal{Y}_2 \quad \Longrightarrow \quad N(\mathbf{P}) \subseteq \mathcal{Y}_1 + \mathcal{Y}_2.$$

Conversely,

$$\begin{aligned}\mathbf{z} \in \mathcal{Y}_1 + \mathcal{Y}_2 &\implies \mathbf{z} = \mathbf{y}_1 + \mathbf{y}_2, \text{ where } \mathbf{y}_i \in \mathcal{Y}_i \text{ for } i = 1, 2 \\ &\implies \mathbf{E}\mathbf{y}_1 = \mathbf{0} \quad \text{and} \quad \mathbf{F}\mathbf{y}_2 = \mathbf{0} \\ &\implies \mathbf{P}\mathbf{z} = 0 \implies \mathcal{Y}_1 + \mathcal{Y}_2 \subseteq N(\mathbf{P}).\end{aligned}$$

Thus $N(\mathbf{P}) = \mathcal{Y}_1 + \mathcal{Y}_2$.

5.9.20. (a) For every inner pseudoinverse, $\mathbf{A}\mathbf{A}^-$ is a projector onto $R(\mathbf{A})$, and $\mathbf{I} - \mathbf{A}^-\mathbf{A}$ is a projector onto $N(\mathbf{A})$. The system being consistent means that

$$\mathbf{b} \in R(\mathbf{A}) = R\left(\mathbf{A}\mathbf{A}^-\right) \implies \mathbf{A}\mathbf{A}^-\mathbf{b} = \mathbf{b},$$

so $\mathbf{A}^-\mathbf{b}$ is a particular solution. Therefore, the general solution of the system is

$$\mathbf{A}^-\mathbf{b} + N(\mathbf{A}) = \mathbf{A}^-\mathbf{b} + R\left(\mathbf{I} - \mathbf{A}^-\mathbf{A}\right).$$

(b) $\mathbf{A}^-\mathbf{A}$ is a projector along $N(\mathbf{A})$, so Exercise 5.9.12 insures $\mathbf{Q}(\mathbf{A}^-\mathbf{A}) = \mathbf{Q}$ and $(\mathbf{A}^-\mathbf{A})\mathbf{Q} = (\mathbf{A}^-\mathbf{A})$. This together with the fact that $\mathbf{P}\mathbf{A} = \mathbf{A}$ allows us to write

$$\mathbf{A}\mathbf{X}\mathbf{A} = \mathbf{A}\mathbf{Q}\mathbf{A}^-\mathbf{P}\mathbf{A} = \mathbf{A}\mathbf{Q}\mathbf{A}^-\mathbf{A} = \mathbf{A}\mathbf{Q} = \mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{Q} = \mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}.$$

Similarly,

$$\begin{aligned}\mathbf{X}\mathbf{A}\mathbf{X} &= (\mathbf{Q}\mathbf{A}^-\mathbf{P})\mathbf{A}(\mathbf{Q}\mathbf{A}^-\mathbf{P}) = \mathbf{Q}\mathbf{A}^-(\mathbf{P}\mathbf{A})\mathbf{Q}\mathbf{A}^-\mathbf{P} = \mathbf{Q}\mathbf{A}^-\mathbf{A}\mathbf{Q}\mathbf{A}^-\mathbf{P} \\ &= \mathbf{Q}(\mathbf{A}^-\mathbf{A}\mathbf{Q})\mathbf{A}^-\mathbf{P} = \mathbf{Q}(\mathbf{A}^-\mathbf{A})\mathbf{A}^-\mathbf{P} = \mathbf{Q}\mathbf{A}^-\mathbf{P} = \mathbf{X},\end{aligned}$$

so $\mathbf{X}$ is a reflexive pseudoinverse for $\mathbf{A}$. To show $\mathbf{X}$ has the prescribed range and nullspace, use the fact that $\mathbf{X}\mathbf{A}$ is a projector onto $R(\mathbf{X})$ and $\mathbf{A}\mathbf{X}$ is a projector along $N(\mathbf{X})$ to write

$$R(\mathbf{X}) = R(\mathbf{X}\mathbf{A}) = R\left(\mathbf{Q}\mathbf{A}^-\mathbf{P}\mathbf{A}\right) = R\left(\mathbf{Q}\mathbf{A}^-\mathbf{A}\right) = R(\mathbf{Q}) = \mathcal{L}$$

and

$$\begin{aligned}N(\mathbf{X}) &= N(\mathbf{A}\mathbf{X}) = N\left(\mathbf{A}\mathbf{Q}\mathbf{A}^-\mathbf{P}\right) = N\left(\mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{Q}\mathbf{A}^-\mathbf{P}\right) \\ &= N\left(\mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{A}^-\mathbf{P}\right) = N\left(\mathbf{A}\mathbf{A}^-\mathbf{P}\right) = N(\mathbf{P}) = \mathcal{M}.\end{aligned}$$

To prove uniqueness, suppose that $\mathbf{X}_1$ and $\mathbf{X}_2$ both satisfy the specified conditions. Then

$$N(\mathbf{X}_2) = \mathcal{M} = R(\mathbf{I} - \mathbf{A}\mathbf{X}_1) \implies \mathbf{X}_2(\mathbf{I} - \mathbf{A}\mathbf{X}_1) = \mathbf{0} \implies \mathbf{X}_2 = \mathbf{X}_2\mathbf{A}\mathbf{X}_1$$

and

$$R(\mathbf{X}_2\mathbf{A}) = R(\mathbf{X}_2) = \mathcal{L} = R(\mathbf{X}_1) \implies \mathbf{X}_2\mathbf{A}\mathbf{X}_1 = \mathbf{X}_1,$$

so $\mathbf{X}_2 = \mathbf{X}_1$.

Solutions for exercises in section 5. 10

5.10.1. Since $index(\mathbf{A}) = k$, we must have that

$$rank\left(\mathbf{A}^{k-1}\right) > rank\left(\mathbf{A}^{k}\right) = rank\left(\mathbf{A}^{k+1}\right) = \cdots = rank\left(\mathbf{A}^{2k}\right) = \cdots,$$

so $rank\left(\mathbf{A}^{k}\right) = rank(\mathbf{A}^{k})^2$, and hence $index(\mathbf{A}^{k}) \leq 1$. But $\mathbf{A}^{k}$ is singular (because $\mathbf{A}$ is singular) so that $index(\mathbf{A}^{k}) > 0$. Consequently, $index(\mathbf{A}^{k}) = 1$.

5.10.2. In this case, $R\left(\mathbf{A}^{k}\right) = \mathbf{0}$ and $N\left(\mathbf{A}^{k}\right) = \Re^n$. The nonsingular component $\mathbf{C}$ in (5.10.5) is missing, and you can take $\mathbf{Q} = \mathbf{I}$, thereby making $\mathbf{A}$ its own core-nilpotent decomposition.

5.10.3. If $\mathbf{A}$ is nonsingular, then $index(\mathbf{A}) = 0$, regardless of whether or not $\mathbf{A}$ is symmetric. If $\mathbf{A}$ is singular and symmetric, we want to prove $index(\mathbf{A}) = 1$. The strategy is to show that $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$ because this implies that $R(\mathbf{A}) \oplus N(\mathbf{A}) = \Re^n$. To do so, start with

$$\mathbf{x} \in R(\mathbf{A}) \cap N(\mathbf{A}) \implies \mathbf{A}\mathbf{x} = \mathbf{0} \quad \text{and} \quad \mathbf{x} = \mathbf{A}\mathbf{y} \quad \text{for some} \quad \mathbf{y}.$$

Now combine this with the symmetry of $\mathbf{A}$ to obtain

$$\mathbf{x}^T = \mathbf{y}^T\mathbf{A}^T = \mathbf{y}^T\mathbf{A} \implies \mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{x} = 0 \implies \|\mathbf{x}\|_2^2 = 0 \implies \mathbf{x} = \mathbf{0}.$$

5.10.4. $index(\mathbf{A}) = 0$ when $\mathbf{A}$ is nonsingular. If $\mathbf{A}$ is singular and normal we want to prove $index(\mathbf{A}) = 1$. The strategy is to show that $R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$ because this implies that $R(\mathbf{A}) \oplus N(\mathbf{A}) = \mathcal{C}^n$. Recall from (4.5.6) that $N(\mathbf{A}) = N(\mathbf{A}^*\mathbf{A})$ and $N(\mathbf{A}^*) = N(\mathbf{A}\mathbf{A}^*)$, so $N(\mathbf{A}) = N(\mathbf{A}^*)$. Start with

$$\mathbf{x} \in R(\mathbf{A}) \cap N(\mathbf{A}) \implies \mathbf{A}\mathbf{x} = \mathbf{0} \quad \text{and} \quad \mathbf{x} = \mathbf{A}\mathbf{y} \quad \text{for some} \quad \mathbf{y},$$

and combine this with $N(\mathbf{A}) = N(\mathbf{A}^*)$ to obtain

$$\mathbf{A}^*\mathbf{x} = \mathbf{0} \text{ and } \mathbf{x} = \mathbf{A}\mathbf{y} \implies \mathbf{x}^*\mathbf{x} = \mathbf{y}^*\mathbf{A}^*\mathbf{x} = 0 \implies \|\mathbf{x}\|_2^2 = 0 \implies \mathbf{x} = \mathbf{0}.$$

5.10.5. Compute $rank\left(\mathbf{A}^0\right) = 3$, $rank(\mathbf{A}) = 2$, $rank\left(\mathbf{A}^2\right) = 1$, and $rank\left(\mathbf{A}^3\right) = 1$, to see that $k = 2$ is the smallest integer such that $rank\left(\mathbf{A}^{k}\right) = rank\left(\mathbf{A}^{k+1}\right)$, so $index(\mathbf{A}) = 2$. The matrix $\mathbf{Q} = [\mathbf{X} \,|\, \mathbf{Y}]$ is a matrix in which the columns of $\mathbf{X}$ are a basis for $R\left(\mathbf{A}^2\right)$, and the columns of $\mathbf{Y}$ are a basis for $N\left(\mathbf{A}^2\right)$. Since

$$\mathbf{E}_{\mathbf{A}^2} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

we have

$$\mathbf{X} = \begin{pmatrix} -8 \\ 12 \\ 8 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} -1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{so} \quad \mathbf{Q} = \begin{pmatrix} -8 & -1 & 0 \\ 12 & 1 & 0 \\ 8 & 0 & 1 \end{pmatrix}.$$

It can now be verified that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{pmatrix} 1/4 & 1/4 & 0 \\ -3 & -2 & 0 \\ -2 & -2 & 1 \end{pmatrix} \begin{pmatrix} -2 & 0 & -4 \\ 4 & 2 & 4 \\ 3 & 2 & 2 \end{pmatrix} \begin{pmatrix} -8 & -1 & 0 \\ 12 & 1 & 0 \\ 8 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 4 \\ 0 & -1 & 2 \end{pmatrix},$$

where

$$\mathbf{C} = [2] \quad \text{and} \quad \mathbf{N} = \begin{pmatrix} -2 & 4 \\ -1 & 2 \end{pmatrix},$$

and $\mathbf{N}^2 = \mathbf{0}$. Finally, $\mathbf{A}^D = \mathbf{Q}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1} = \begin{pmatrix} -1 & -1 & 0 \\ 3/2 & 3/2 & 0 \\ 1 & 1 & 0 \end{pmatrix}$.

5.10.6. (a) Because

$$\mathbf{J} - \lambda\mathbf{I} = \begin{pmatrix} 1-\lambda & 0 & 0 & 0 & 0 \\ 0 & 1-\lambda & 0 & 0 & 0 \\ 0 & 0 & 1-\lambda & 0 & 0 \\ 0 & 0 & 0 & 2-\lambda & 0 \\ 0 & 0 & 0 & 0 & 2-\lambda \end{pmatrix},$$

and because a diagonal matrix is singular if and only if it has a zero-diagonal entry, it follows that $\mathbf{J} - \lambda\mathbf{I}$ is singular if and only if $\lambda = 1$ or $\lambda = 2$, so $\lambda_1 = 1$ and $\lambda_2 = 2$ are the two eigenvalues of $\mathbf{J}$. To find the index of λ_1, use block multiplication to observe that

$$\mathbf{J} - \mathbf{I} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{2\times 2} \end{pmatrix} \implies rank\,(\mathbf{J} - \mathbf{I}) = 2 = rank\,(\mathbf{J} - \mathbf{I})^2.$$

Therefore, $index(\lambda_1) = 1$. Similarly,

$$\mathbf{J} - 2\mathbf{I} = \begin{pmatrix} -\mathbf{I}_{3\times 3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad rank\,(\mathbf{J} - 2\mathbf{I}) = 3 = rank\,(\mathbf{J} - 2\mathbf{I})^2,$$

so $index(\lambda_2) = 1$.

(b) Since

$$\mathbf{J} - \lambda\mathbf{I} = \begin{pmatrix} 1-\lambda & 1 & 0 & 0 & 0 \\ 0 & 1-\lambda & 1 & 0 & 0 \\ 0 & 0 & 1-\lambda & 0 & 0 \\ 0 & 0 & 0 & 2-\lambda & 1 \\ 0 & 0 & 0 & 0 & 2-\lambda \end{pmatrix},$$

and since a triangular matrix is singular if and only if there exists a zero-diagonal entry (i.e., a zero pivot), it follows that $\mathbf{J} - \lambda\mathbf{I}$ is singular if and only if $\lambda = 1$

or $\lambda = 2$, so $\lambda_1 = 1$ and $\lambda_2 = 2$ are the two eigenvalues of $\mathbf{J}$. To find the index of λ_1, use block multiplication to compute

$$\mathbf{J}-\mathbf{I} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad (\mathbf{J}-\mathbf{I})^2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$(\mathbf{J}-\mathbf{I})^3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad (\mathbf{J}-\mathbf{I})^4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since

$$rank\,(\mathbf{J}-\mathbf{I}) > rank\,(\mathbf{J}-\mathbf{I})^2 > rank\,(\mathbf{J}-\mathbf{I})^3 = rank\,(\mathbf{J}-\mathbf{I})^4,$$

it follows that $index(\lambda_1) = 3$. A similar computation using λ_2 shows that

$$rank\,(\mathbf{J}-2\mathbf{I}) > rank\,(\mathbf{J}-2\mathbf{I})^2 = rank\,(\mathbf{J}-2\mathbf{I})^3,$$

so $index(\lambda_2) = 2$. The fact that eigenvalues associated with diagonal matrices have index 1 while eigenvalues associated with triangular matrices can have higher indices is no accident. This will be discussed in detail in §7.8 (p. 587).

5.10.7. (a) If $\mathbf{P}$ is a projector, then, by (5.9.13), $\mathbf{P} = \mathbf{P}^2$, so $rank\,(\mathbf{P}) = rank\,(\mathbf{P}^2)$, and hence $index(\mathbf{P}) \le 1$. If $\mathbf{P} \ne \mathbf{I}$, then $\mathbf{P}$ is singular, and thus $index(\mathbf{P}) = 1$. If $\mathbf{P} = \mathbf{I}$, then $index(\mathbf{P}) = 0$. An alternate argument could be given on the basis of the observation that $\Re^n = R\,(\mathbf{P}) \oplus N\,(\mathbf{P})$.

(b) Recall from (5.9.12) that if the columns of $\mathbf{X}$ and $\mathbf{Y}$ constitute bases for $R\,(\mathbf{P})$ and $N\,(\mathbf{P})$, respectively, then for $\mathbf{Q} = \big[\mathbf{X}\,|\,\mathbf{Y}\big]$,

$$\mathbf{Q}^{-1}\mathbf{P}\mathbf{Q} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and it follows that $\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ is the core-nilpotent decomposition for $\mathbf{P}$.

5.10.8. Suppose that $\sum_{i=0}^{k-1} \alpha_i \mathbf{N}^i \mathbf{x} = \mathbf{0}$, and multiply both sides by $\mathbf{N}^{k-1}$ to obtain $\alpha_0 \mathbf{N}^{k-1}\mathbf{x} = \mathbf{0}$. By assumption, $\mathbf{N}^{k-1}\mathbf{x} \ne \mathbf{0}$, so $\alpha_0 = 0$, and hence $\sum_{i=1}^{k-1} \alpha_i \mathbf{N}^i \mathbf{x} = \mathbf{0}$. Now multiply both sides of this equation by $\mathbf{N}^{k-2}$ to produce $\alpha_1 \mathbf{N}^{k-1}\mathbf{x} = \mathbf{0}$, and conclude that $\alpha_1 = 0$. Continuing in this manner (or by making a formal induction argument) gives $\alpha_0 = \alpha_1 = \alpha_2 = \cdots = \alpha_{k-1} = 0$.

5.10.9. (a) $\mathbf{b} \in R\,(\mathbf{A}^k) \subseteq R\,(\mathbf{A}) \implies \mathbf{b} \in R\,(\mathbf{A}) \implies \mathbf{A}\mathbf{x} = \mathbf{b}$ is consistent.

(b) We saw in Example 5.10.5 that when considered as linear operators restricted to $R(\mathbf{A}^k)$, both $\mathbf{A}$ and $\mathbf{A}^D$ are invertible, and in fact they are true inverses of each other. Consequently, $\mathbf{A}$ and $\mathbf{A}^D$ are one-to-one mappings on $R(\mathbf{A}^k)$ (recall Exercise 4.7.18), so for each $\mathbf{b} \in R(\mathbf{A}^k)$ there is a unique $\mathbf{x} \in R(\mathbf{A}^k)$ such that $\mathbf{Ax} = \mathbf{b}$, and this unique $\mathbf{x}$ is given by $\mathbf{x} = \left(\mathbf{A}/_{R(\mathbf{A}^k)}\right)^{-1}\mathbf{b} = \mathbf{A}^D\mathbf{b}$.

(c) Part (b) shows that $\mathbf{A}^D\mathbf{b}$ is a particular solution. The desired result follows because the general solution is any particular solution plus the general solution of the associated homogeneous equation.

5.10.10. Notice that $\mathbf{AA}^D = \mathbf{Q}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1}$, and use the results from Example 5.10.3 (p. 398). $\mathbf{I} - \mathbf{AA}^D$ is the complementary projector, so it projects onto $N(\mathbf{A}^k)$ along $R(\mathbf{A}^k)$.

5.10.11. In each case verify that the axioms (A1), (A2), (A4), and (A5) in the definition of a vector space given on p. 160 hold for matrix multiplication (rather than +). In parts (a) and (b) the identity element is the ordinary identity matrix, and the inverse of each member is the ordinary inverse. In part (c), the identity element is $\mathbf{E} = \begin{pmatrix}1/2 & 1/2\\ 1/2 & 1/2\end{pmatrix}$ because $\mathbf{AE} = \mathbf{A} = \mathbf{EA}$ for each $\mathbf{A} \in \mathcal{G}$, and $\begin{pmatrix}\alpha & \alpha\\ \alpha & \alpha\end{pmatrix}^{\#} = \dfrac{1}{4\alpha}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$ because $\mathbf{AA}^{\#} = \mathbf{E} = \mathbf{A}^{\#}\mathbf{A}$.

5.10.12. (a) $\Longrightarrow$ (b) : If $\mathbf{A}$ belongs to a matrix group $\mathcal{G}$ in which the identity element is $\mathbf{E}$, and if $\mathbf{A}^{\#}$ is the inverse of $\mathbf{A}$ in $\mathcal{G}$, then $\mathbf{A}^{\#}\mathbf{A}^2 = \mathbf{EA} = \mathbf{A}$, so

$$\begin{aligned}\mathbf{x} \in R(\mathbf{A}) \cap N(\mathbf{A}) &\Longrightarrow \mathbf{x} = \mathbf{Ay} \text{ for some } \mathbf{y} \text{ and } \mathbf{Ax} = \mathbf{0}\\ &\Longrightarrow \mathbf{Ay} = \mathbf{A}^{\#}\mathbf{A}^2\mathbf{y} = \mathbf{A}^{\#}\mathbf{Ax} = \mathbf{0}\\ &\Longrightarrow \mathbf{x} = \mathbf{0}.\end{aligned}$$

(b) $\Longrightarrow$ (c) : Suppose $\mathbf{A}$ is $n \times n$, and let $\mathcal{B}_R$ and $\mathcal{B}_N$ be bases for $R(\mathbf{A})$ and $N(\mathbf{A})$, respectively. Verify that $\mathcal{B} = R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0}$ implies $\mathcal{B}_R \cap \mathcal{B}_N$ is a linearly independent set, and use the fact that there are n vectors in $\mathcal{B}$ to conclude that $\mathcal{B}$ is a basis for $\Re^n$. Statement (c) now follows from (5.9.4).

(c) $\Longrightarrow$ (d) : Use the fact that $R(\mathbf{A}^k) \cap N(\mathbf{A}^k) = \mathbf{0}$.

(d) $\Longrightarrow$ (e) : Use the result of Example 5.10.5 together with the fact that the only nilpotent matrix of index 1 is the zero matrix.

(e) $\Longrightarrow$ (a) : It is straightforward to verify that the set

$$\mathcal{G} = \left\{\mathbf{Q}\begin{pmatrix}\mathbf{X}_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Q}^{-1} \,\middle|\, \mathbf{X} \text{ is nonsingular}\right\}$$

is a matrix group, and it's clear that $\mathbf{A} \in \mathcal{G}$.

5.10.13. (a) Use part (e) of Exercise 5.10.12 to write $\mathbf{A} = \mathbf{Q}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}^{-1}$. For the given $\mathbf{E}$, verify that $\mathbf{EA} = \mathbf{AE} = \mathbf{A}$ for all $\mathbf{A} \in \mathcal{G}$. The fact that $\mathbf{E}$ is the desired projector follows from (5.9.12).

(b) Simply verify that $\mathbf{AA}^{\#} = \mathbf{A}^{\#}\mathbf{A} = \mathbf{E}$. Notice that the group inverse agrees with the Drazin inverse of $\mathbf{A}$ described in Example 5.10.5. However, the Drazin inverse exists for all square matrices, but the concept of a group inverse makes sense only for group matrices—i.e., when $index(\mathbf{A}) = 1$.

Solutions for exercises in section 5. 11

5.11.1. Proceed as described on p. 199 to determine the following bases for each of the four fundamental subspaces.

$$R(\mathbf{A}) = span\left\{\begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}\right\} \quad N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\right\}$$

$$N(\mathbf{A}) = span\left\{\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}\right\} \qquad R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}\right\}$$

Since each vector in a basis for $R(\mathbf{A})$ is orthogonal to each vector in a basis for $N\left(\mathbf{A}^T\right)$, it follows that $R(\mathbf{A}) \perp N\left(\mathbf{A}^T\right)$. The same logic also explains why $N(\mathbf{A}) \perp R\left(\mathbf{A}^T\right)$. Notice that $R(\mathbf{A})$ is a plane through the origin in $\Re^3$, and $N\left(\mathbf{A}^T\right)$ is the line through the origin perpendicular to this plane, so it is evident from the parallelogram law that $R(\mathbf{A}) \oplus N\left(\mathbf{A}^T\right) = \Re^3$. Similarly, $N(\mathbf{A})$ is the line through the origin normal to the plane defined by $R\left(\mathbf{A}^T\right)$, so $N(\mathbf{A}) \oplus R\left(\mathbf{A}^T\right) = \Re^3$.

5.11.2. $\mathcal{V}^{\perp} = \mathbf{0}$, and $\mathbf{0}^{\perp} = \mathcal{V}$.

5.11.3. If $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 0 & 1 \\ 3 & 6 \end{pmatrix}$, then $R(\mathbf{A}) = \mathcal{M}$, so (5.11.5) insures $\mathcal{M}^{\perp} = N\left(\mathbf{A}^T\right)$. Using row operations, a basis for $N\left(\mathbf{A}^T\right)$ is computed to be $\left\{\begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ 0 \\ 0 \\ 1 \end{pmatrix}\right\}$.

5.11.4. Verify that $\mathcal{M}^{\perp}$ is closed with respect to vector addition and scalar multiplication. If $\mathbf{x}, \mathbf{y} \in \mathcal{M}^{\perp}$, then $\langle \mathbf{m}|\mathbf{x}\rangle = 0 = \langle \mathbf{m}|\mathbf{y}\rangle$ for each $\mathbf{m} \in \mathcal{M}$ so that $\langle \mathbf{m}|\mathbf{x}+\mathbf{y}\rangle = 0$ for each $\mathbf{m} \in \mathcal{M}$, and thus $\mathbf{x}+\mathbf{y} \in \mathcal{M}^{\perp}$. Similarly, for every scalar α we have $\langle \mathbf{m}|\alpha\mathbf{x}\rangle = \alpha\langle \mathbf{m}|\mathbf{x}\rangle = 0$ for each $\mathbf{m} \in \mathcal{M}$, so $\alpha\mathbf{x} \in \mathcal{M}^{\perp}$.

5.11.5. (a) $\mathbf{x} \in \mathcal{N}^{\perp} \Longrightarrow \mathbf{x} \perp \mathcal{N} \supseteq \mathcal{M} \Longrightarrow \mathbf{x} \perp \mathcal{M} \Longrightarrow \mathbf{x} \in \mathcal{M}^{\perp}$.

(b) Simply observe that

$$\begin{aligned}\mathbf{x} \in (\mathcal{M}+\mathcal{N})^{\perp} &\iff \mathbf{x} \perp (\mathcal{M}+\mathcal{N}) \\ &\iff \mathbf{x} \perp \mathcal{M} \text{ and } \mathbf{x} \perp \mathcal{N} \\ &\iff \mathbf{x} \in \left(\mathcal{M}^{\perp} \cap \mathcal{N}^{\perp}\right).\end{aligned}$$

(c) Use part (b) together with (5.11.4) to write

$$\left(\mathcal{M}^{\perp}+\mathcal{N}^{\perp}\right)^{\perp} = \mathcal{M}^{\perp^{\perp}} \cap \mathcal{N}^{\perp^{\perp}} = \mathcal{M} \cap \mathcal{N},$$

and then perp both sides.

5.11.6. Use the fact that $\dim R\left(\mathbf{A}^T\right) = rank\left(\mathbf{A}^T\right) = rank\left(\mathbf{A}\right) = \dim R\left(\mathbf{A}\right)$ together with (5.11.7) to conclude that

$$n = \dim N\left(\mathbf{A}\right) + \dim R\left(\mathbf{A}^T\right) = \dim N\left(\mathbf{A}\right) + \dim R\left(\mathbf{A}\right).$$

5.11.7. $\mathbf{U}$ is a unitary matrix in which the columns of $\mathbf{U}_1$ are an orthonormal basis for $R\left(\mathbf{A}\right)$ and the columns of $\mathbf{U}_2$ are an orthonormal basis for $N\left(\mathbf{A}^T\right)$, so setting $\mathbf{X} = \mathbf{U}_1$, $\mathbf{Y} = \mathbf{U}_2$, and $\left[\mathbf{X}\,|\,\mathbf{Y}\right]^{-1} = \mathbf{U}^T$ in (5.9.12) produces $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T$. According to (5.9.9), the projector onto $N\left(\mathbf{A}^T\right)$ along $R\left(\mathbf{A}\right)$ is $\mathbf{I}-\mathbf{P} = \mathbf{I} - \mathbf{U}_1\mathbf{U}_1^T = \mathbf{U}_2\mathbf{U}_2^T$.

5.11.8. Start with the first column of $\mathbf{A}$, and set $\mathbf{u} = \mathbf{A}_{*1} + 6\mathbf{e}_1 = \begin{pmatrix}2 & 2 & -4\end{pmatrix}^T$ to obtain

$$\mathbf{R}_1 = \mathbf{I} - \frac{2\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{3}\begin{pmatrix} 2 & -1 & 2 \\ -1 & 2 & 2 \\ 2 & 2 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_1\mathbf{A} = \begin{pmatrix} -6 & 0 & -6 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 \end{pmatrix}.$$

Now set $\mathbf{u} = \begin{pmatrix} 0 \\ -3 \end{pmatrix} + 3\mathbf{e}_1 = \begin{pmatrix} 3 \\ -3 \end{pmatrix}$ to get

$$\hat{\mathbf{R}}_2 = \mathbf{I} - \frac{2\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_2 = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

so

$$\mathbf{P} = \mathbf{R}_2\mathbf{R}_1 = \frac{1}{3}\begin{pmatrix} 2 & -1 & 2 \\ 2 & 2 & -1 \\ -1 & 2 & 2 \end{pmatrix} \text{ and } \mathbf{PA} = \begin{pmatrix} -6 & 0 & -6 & -3 \\ 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{B} \\ \mathbf{0} \end{pmatrix}.$$

Therefore, $rank\,(\mathbf{A}) = 2$, and orthonormal bases for $R\,(\mathbf{A})$ and $N\left(\mathbf{A}^T\right)$ are extracted from the columns of $\mathbf{U} = \mathbf{P}^T$ as shown below.

$$R\,(\mathbf{A}) = span\left\{\begin{pmatrix} 2/3 \\ -1/3 \\ 2/3 \end{pmatrix}, \begin{pmatrix} 2/3 \\ 2/3 \\ -1/3 \end{pmatrix}\right\} \quad \text{and} \quad N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} -1/3 \\ 2/3 \\ 2/3 \end{pmatrix}\right\}$$

Now work with $\mathbf{B}^T$, and set $\mathbf{u} = (\mathbf{B}_{1*})^T + 9\mathbf{e}_1 = \begin{pmatrix} 3 & 0 & -6 & -3 \end{pmatrix}^T$ to get

$$\mathbf{Q} = \mathbf{I} - \frac{2\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{3}\begin{pmatrix} 2 & 0 & 2 & 1 \\ 0 & 3 & 0 & 0 \\ 2 & 0 & -1 & -2 \\ 1 & 0 & -2 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{Q}\mathbf{B}^T = \begin{pmatrix} -9 & 0 \\ 0 & -3 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{T} \\ \mathbf{0} \end{pmatrix}.$$

Orthonormal bases for $R\left(\mathbf{A}^T\right)$ and $N\,(\mathbf{A})$ are extracted from the columns of $\mathbf{V} = \mathbf{Q}^T = \mathbf{Q}$ as shown below.

$$R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} 2/3 \\ 0 \\ 2/3 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}\right\} \text{ and } N\,(\mathbf{A}) = span\left\{\begin{pmatrix} 2/3 \\ 0 \\ -1/3 \\ -2/3 \end{pmatrix}, \begin{pmatrix} 1/3 \\ 0 \\ -2/3 \\ 2/3 \end{pmatrix}\right\}$$

A URV factorization is obtained by setting $\mathbf{U} = \mathbf{P}^T$, $\mathbf{V} = \mathbf{Q}^T$, and

$$\mathbf{R} = \begin{pmatrix} \mathbf{T}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} -9 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

5.11.9. Using $\mathbf{E_A} = \begin{pmatrix} 1 & 0 & 1 & 1/2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ along with the standard methods of Chapter 4, we have

$$R\,(\mathbf{A}) = span\left\{\begin{pmatrix} -4 \\ 2 \\ -4 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}\right\} \quad \text{and} \quad N\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}\right\},$$

$$R\left(\mathbf{A}^T\right) = span\left\{\begin{pmatrix} 1 \\ 0 \\ 1 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}\right\} \text{ and } N\,(\mathbf{A}) = span\left\{\begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1/2 \\ 0 \\ 0 \\ 1 \end{pmatrix}\right\}.$$

Applying the Gram–Schmidt procedure to each of these sets produces the following orthonormal bases for the four fundamental subspaces.

$$\mathcal{B}_{R(\mathbf{A})} = \left\{\frac{1}{3}\begin{pmatrix} -2 \\ 1 \\ -2 \end{pmatrix}, \frac{1}{3}\begin{pmatrix} -2 \\ -2 \\ 1 \end{pmatrix}\right\} \qquad \mathcal{B}_{N\left(\mathbf{A}^T\right)} = \left\{\frac{1}{3}\begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}\right\}$$

$$\mathcal{B}_{R(\mathbf{A}^T)} = \left\{ \frac{1}{3}\begin{pmatrix} 2 \\ 0 \\ 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \right\} \qquad \mathcal{B}_{N(\mathbf{A})} = \left\{ \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{3\sqrt{2}}\begin{pmatrix} -1 \\ 0 \\ -1 \\ 4 \end{pmatrix} \right\}$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ were defined in (5.11.8) to be

$$\mathbf{U} = \left(\mathcal{B}_{R(\mathbf{A})} \cup \mathcal{B}_{N(\mathbf{A}^T)}\right) = \frac{1}{3}\begin{pmatrix} -2 & -2 & -1 \\ 1 & -2 & 2 \\ -2 & 1 & 2 \end{pmatrix}$$

and

$$\mathbf{V} = \left(\mathcal{B}_{R(\mathbf{A}^T)} \cup \mathcal{B}_{N(\mathbf{A})}\right) = \frac{1}{3}\begin{pmatrix} 2 & 0 & -3/\sqrt{2} & -1/\sqrt{2} \\ 0 & 3 & 0 & 0 \\ 2 & 0 & 3/\sqrt{2} & -1/\sqrt{2} \\ 1 & 0 & 0 & 4/\sqrt{2} \end{pmatrix}.$$

Direct multiplication now produces

$$\mathbf{R} = \mathbf{U}^T\mathbf{A}\mathbf{V} = \begin{pmatrix} 9 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

5.11.10. According to the discussion of projectors on p. 386, the *unique* vectors satisfying $\mathbf{v} = \mathbf{x}+\mathbf{y}$, $\mathbf{x} \in R(\mathbf{A})$, and $\mathbf{y} \in N(\mathbf{A}^T)$ are given by $\mathbf{x} = \mathbf{P}\mathbf{v}$ and $\mathbf{y} = (\mathbf{I}-\mathbf{P})\mathbf{v}$, where $\mathbf{P}$ is the projector onto $R(\mathbf{A})$ along $N(\mathbf{A}^T)$. Use the results of Exercise 5.11.7 and Exercise 5.11.8 to compute

$$\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T = \frac{1}{9}\begin{pmatrix} 8 & 2 & 2 \\ 2 & 5 & -4 \\ 2 & -4 & 5 \end{pmatrix}, \mathbf{x} = \mathbf{P}\mathbf{v} = \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix}, \mathbf{y} = (\mathbf{I}-\mathbf{P})\mathbf{v} = \begin{pmatrix} -1 \\ 2 \\ 2 \end{pmatrix}.$$

5.11.11. Observe that

$$\begin{aligned} R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0} &\implies \mathit{index}(\mathbf{A}) \leq 1, \\ R(\mathbf{A}) \not\perp N(\mathbf{A}) &\implies \mathbf{A} \text{ is singular}, \\ R(\mathbf{A}) \not\perp N(\mathbf{A}) &\implies R(\mathbf{A}^T) \neq R(\mathbf{A}). \end{aligned}$$

It is now trial and error to build a matrix that satisfies the three conditions on the right-hand side. One such matrix is $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$.

5.11.12. $R(\mathbf{A}) \perp N(\mathbf{A}) \implies R(\mathbf{A}) \cap N(\mathbf{A}) = \mathbf{0} \implies \mathit{index}(\mathbf{A}) = 1$ by using (5.10.4). The example in the solution to Exercise 5.11.11 shows that the converse is false.

5.11.13. The facts that real symmetric $\implies$ hermitian $\implies$ normal are direct consequences of the definitions. To show that normal $\implies$ RPN, use (4.5.5) to write $R(\mathbf{A}) = R(\mathbf{A}\mathbf{A}^*) = R(\mathbf{A}^*\mathbf{A}) = R(\mathbf{A}^*)$. The matrix $\begin{pmatrix} 1 & \mathrm{i} \\ -\mathrm{i} & 2 \end{pmatrix}$ is hermitian but not symmetric. To construct a matrix that is normal but not hermitian or

real symmetric, try to find an example with real numbers. If $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then

$$\mathbf{A}\mathbf{A}^T = \begin{pmatrix} a^2+b^2 & ac+bd \\ ac+bd & c^2+d^2 \end{pmatrix} \quad \text{and} \quad \mathbf{A}^T\mathbf{A} = \begin{pmatrix} a^2+c^2 & ab+cd \\ ab+cd & b^2+d^2 \end{pmatrix},$$

so we need to have $b^2 = c^2$. One such matrix is $\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$. To construct a singular matrix that is RPN but not normal, try again to find an example with real numbers. For any orthogonal matrix $\mathbf{P}$ and nonsingular matrix $\mathbf{C}$, the matrix $\mathbf{A} = \mathbf{P}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^T$ is RPN. To prevent $\mathbf{A}$ from being normal, simply choose $\mathbf{C}$ to be nonnormal. For example, let $\mathbf{C} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $\mathbf{P} = \mathbf{I}$.

5.11.14. (a) $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* \implies (\mathbf{A}-\lambda\mathbf{I})^*(\mathbf{A}-\lambda\mathbf{I}) = (\mathbf{A}-\lambda\mathbf{I})(\mathbf{A}-\lambda\mathbf{I})^* \implies (\mathbf{A}-\lambda\mathbf{I})$ is normal $\implies (\mathbf{A}-\lambda\mathbf{I})$ is RPN $\implies R(\mathbf{A}-\lambda\mathbf{I}) \perp N(\mathbf{A}-\lambda\mathbf{I})$.
(b) Suppose $\mathbf{x} \in N(\mathbf{A}-\lambda\mathbf{I})$ and $\mathbf{y} \in N(\mathbf{A}-\mu\mathbf{I})$, and use the fact that $N(\mathbf{A}-\lambda\mathbf{I}) = N(\mathbf{A}-\lambda\mathbf{I})^*$ to write

$$\begin{aligned}(\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = \mathbf{0} \implies \mathbf{0} = \mathbf{x}^*(\mathbf{A}-\lambda\mathbf{I}) \implies 0 = \mathbf{x}^*(\mathbf{A}-\lambda\mathbf{I})\mathbf{y} \\ = \mathbf{x}^*(\mu\mathbf{y}-\lambda\mathbf{y}) = \mathbf{x}^*\mathbf{y}(\mu-\lambda) \implies \mathbf{x}^*\mathbf{y} = 0.\end{aligned}$$

Solutions for exercises in section 5. 12

5.12.1. Since $\mathbf{C}^T\mathbf{C} = \begin{pmatrix} 25 & 0 \\ 0 & 100 \end{pmatrix}$, $\sigma_1^2 = 100$, and it's clear that $\mathbf{x} = \mathbf{e}_2$ is a vector such that $(\mathbf{C}^T\mathbf{C} - 100\mathbf{I})\mathbf{x} = 0$ and $\|\mathbf{x}\|_2 = 1$. Let $\mathbf{y} = \mathbf{C}\mathbf{x}/\sigma_1 = \begin{pmatrix} -3/5 \\ -4/5 \end{pmatrix}$. Following the procedure in Example 5.6.3, set $\mathbf{u}_x = \mathbf{x} - \mathbf{e}_1$ and $\mathbf{u}_y = \mathbf{y} - \mathbf{e}_1$, and construct

$$\mathbf{R}_x = \mathbf{I} - 2\frac{\mathbf{u}_x\mathbf{u}_x^T}{\mathbf{u}_x^T\mathbf{u}_x} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_y = \mathbf{I} - 2\frac{\mathbf{u}_y\mathbf{u}_y^T}{\mathbf{u}_y^T\mathbf{u}_y} = \begin{pmatrix} -3/5 & -4/5 \\ -4/5 & 3/5 \end{pmatrix}.$$

Since $\mathbf{R}_y\mathbf{C}\mathbf{R}_x = \begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix} = \mathbf{D}$, it follows that $\mathbf{C} = \mathbf{R}_y\mathbf{D}\mathbf{R}_x$ is a singular value decomposition of $\mathbf{C}$.

5.12.2. $\nu_1^2(\mathbf{A}) = \sigma_1^2 = \|\mathbf{A}\|_2^2$ needs no proof—it's just a restatement of (5.12.4). The fact that $\nu_r^2(\mathbf{A}) = \|\mathbf{A}\|_F^2$ amounts to observing that

$$\|\mathbf{A}\|_F^2 = trace\left(\mathbf{A}^T\mathbf{A}\right) = trace\mathbf{V}\begin{pmatrix} \mathbf{D}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T = trace\left(\mathbf{D^2}\right) = \sigma_1^2 + \cdots + \sigma_r^2.$$

5.12.3. If $\sigma_1 \geq \cdots \geq \sigma_r$ are the nonzero singular values for $\mathbf{A}$, then it follows from Exercise 5.12.2 that $\|\mathbf{A}\|_2^2 = \sigma_1^2 \leq \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2 = \|\mathbf{A}\|_F^2 \leq n\sigma_1^2 = n\|\mathbf{A}\|_2^2$.

5.12.4. If $rank\,(\mathbf{A}+\mathbf{E}) = k < r$, then (5.12.10) implies that

$$\|\mathbf{E}\|_2 = \|\mathbf{A} - (\mathbf{A}+\mathbf{E})\|_2 \geq \min_{rank(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_2 = \sigma_{k+1} \geq \sigma_r,$$

which is impossible. Hence $rank\,(\mathbf{A}+\mathbf{E}) \geq r = rank\,(\mathbf{A})$.

5.12.5. The argument is almost identical to that given for the nonsingular case except that $\mathbf{A}^\dagger$ replaces $\mathbf{A}^{-1}$. Start with SVDs

$$\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T \quad \text{and} \quad \mathbf{A}^\dagger = \mathbf{V}\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T,$$

where $\mathbf{D} = \text{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_r)$, and note that $\left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{x}\right\|_2 \leq \left\|\mathbf{A}^\dagger\mathbf{A}\right\|_2 \|\mathbf{x}\|_2 = 1$ with equality holding when $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$ (i.e., when $r = n$). For each $\mathbf{y} \in \mathbf{A}(\mathcal{S}_2)$ there is an $\mathbf{x} \in \mathcal{S}_2$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$, so, with $\mathbf{w} = \mathbf{U}^T\mathbf{y}$,

$$\begin{aligned} 1 \geq \left\|\mathbf{A}^\dagger\mathbf{A}\mathbf{x}\right\|_2^2 &= \left\|\mathbf{A}^\dagger\mathbf{y}\right\|_2^2 = \left\|\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 = \left\|\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}\right\|_2^2 \\ &= \left\|\mathbf{D}^{-1}\mathbf{w}\right\|_2^2 = \frac{w_1^2}{\sigma_1^2} + \frac{w_2^2}{\sigma_2^2} + \cdots + \frac{w_r^2}{\sigma_r^2} \end{aligned}$$

with equality holding when $r = n$. In other words, the set $\mathbf{U}^T\mathbf{A}(\mathcal{S}_2)$ is an ellipsoid (degenerate if $r < n$) whose k^{th} semiaxis has length σ_k. To resolve the inequality with what it means for points to be *on* an ellipsoid, realize that the surface of a degenerate ellipsoid (one having some semiaxes with zero length) is actually the set of all points *in and on* a smaller dimension ellipsoid. For example, visualize an ellipsoid in $\Re^3$, and consider what happens as one of its semiaxes shrinks to zero. The skin of the three-dimensional ellipsoid degenerates to a solid planar ellipse. In other words, all points *on* a degenerate ellipsoid with semiaxes of length $\sigma_1 \neq 0$, $\sigma_2 \neq 0$, $\sigma_3 = 0$ are actually points *on and inside* a planar ellipse with semiaxes of length σ_1 and σ_2. Arguing that the kth semiaxis of $\mathbf{A}(\mathcal{S}_2)$ is $\sigma_k\mathbf{U}_{*k} = \mathbf{A}\mathbf{V}_{*k}$ is the same as the nonsingular case given in the text.

5.12.6. If $\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T$ and $\mathbf{A}^\dagger_{n\times m} = \mathbf{V}\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T$ are SVDs in which $\mathbf{V} = \left(\mathbf{V}_1 \,|\, \mathbf{V}_2\right)$, then the columns of $\mathbf{V}_1$ are an orthonormal basis for $R\left(\mathbf{A}^T\right)$, so $\mathbf{x} \in R\left(\mathbf{A}^T\right)$ and $\|\mathbf{x}\|_2 = 1$ if and only if $\mathbf{x} = \mathbf{V}_1\mathbf{y}$ with $\|\mathbf{y}\|_2 = 1$. Since the 2-norm is unitarily invariant (Exercise 5.6.9),

$$\min_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in R(\mathbf{A}^T)}} \|\mathbf{A}\mathbf{x}\|_2 = \min_{\|\mathbf{y}\|_2=1} \|\mathbf{A}\mathbf{V}_1\mathbf{y}\|_2 = \min_{\|\mathbf{y}\|_2=1} \|\mathbf{D}\mathbf{y}\|_2 = \frac{1}{\|\mathbf{D}^{-1}\|_2} = \sigma_r = \frac{1}{\|\mathbf{A}^\dagger\|_2}.$$

5.12.7. $\mathbf{x} = \mathbf{A}^\dagger \mathbf{b}$ and $\tilde{\mathbf{x}} = \mathbf{A}^\dagger(\mathbf{b} - \mathbf{e})$ are the respective solutions of minimal 2-norm of $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} = \mathbf{b} - \mathbf{e}$. The development of the more general bound is the same as for (5.12.8).

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{A}^\dagger(\mathbf{b} - \tilde{\mathbf{b}})\| \le \|\mathbf{A}^\dagger\| \, \|\mathbf{b} - \tilde{\mathbf{b}}\|,$$
$$\mathbf{b} = \mathbf{A}\mathbf{x} \implies \|\mathbf{b}\| \le \|\mathbf{A}\| \, \|\mathbf{x}\| \implies 1/\|\mathbf{x}\| \le \|\mathbf{A}\|/\|\mathbf{b}\|,$$

so

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \left(\|\mathbf{A}^\dagger\| \, \|\mathbf{b} - \tilde{\mathbf{b}}\|\right) \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|} = \kappa \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}.$$

Similarly,

$$\|\mathbf{b} - \tilde{\mathbf{b}}\| = \|\mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})\| \le \|\mathbf{A}\| \, \|\mathbf{x} - \tilde{\mathbf{x}}\|,$$
$$\mathbf{x} = \mathbf{A}^\dagger \mathbf{b} \implies \|\mathbf{x}\| \le \|\mathbf{A}^\dagger\| \, \|\mathbf{b}\| \implies 1/\|\mathbf{b}\| \le \|\mathbf{A}^\dagger\|/\|\mathbf{x}\|,$$

so

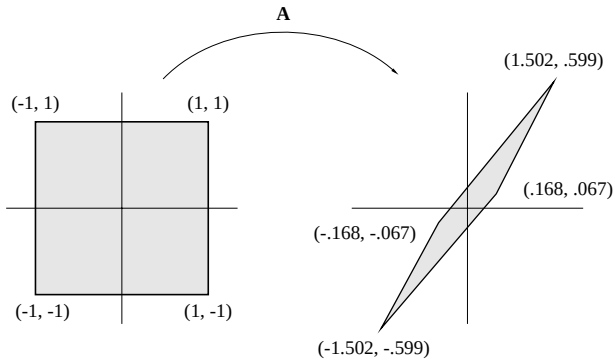
$$\frac{\|\mathbf{b} - \tilde{\mathbf{b}}\|}{\|\mathbf{b}\|} \le (\|\mathbf{A}\| \, \|\mathbf{x} - \tilde{\mathbf{x}}\|) \frac{\|\mathbf{A}^\dagger\|}{\|\mathbf{x}\|} = \kappa \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}.$$

Equality was attained in Example 5.12.1 by choosing $\mathbf{b}$ and $\mathbf{e}$ to point in special directions. But for these choices, $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}} = \mathbf{b} - \mathbf{e}$ cannot be guaranteed to be consistent for all singular or rectangular matrices $\mathbf{A}$, so the answer to the second part is "no." However, the argument of Example 5.12.1 proves equality for all $\mathbf{A}$ such that $\mathbf{A}\mathbf{A}^\dagger = \mathbf{I}$ (i.e., when $rank\,(\mathbf{A}_{m\times n}) = m$).

5.12.8. If $\mathbf{A} = \mathbf{U}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T$ is an SVD, then $\mathbf{A}^T\mathbf{A} + \epsilon\mathbf{I} = \mathbf{U}\begin{pmatrix} \mathbf{D}^2 + \epsilon\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \epsilon\mathbf{I} \end{pmatrix}\mathbf{V}^T$ is an SVD with no zero singular values, so it's nonsingular. Furthermore,

$$(\mathbf{A}^T\mathbf{A} + \epsilon\mathbf{I})^{-1}\mathbf{A}^T = \mathbf{U}\begin{pmatrix} (\mathbf{D}^2 + \epsilon\mathbf{I})^{-1}\mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T \to \mathbf{U}\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T = \mathbf{A}^\dagger.$$

5.12.9. Since $\mathbf{A}^{-1} = \begin{pmatrix} -266000 & 667000 \\ 333000 & -835000 \end{pmatrix}$, $\kappa_\infty = \|\mathbf{A}\|_\infty \left\|\mathbf{A}^{-1}\right\|_\infty = 1,754,336.$ Similar to the 2-norm situation discussed in Example 5.12.1, the worst case is realized when $\mathbf{b}$ is in the direction of a maximal vector in $\mathbf{A}(\mathcal{S}_\infty)$ while $\mathbf{e}$ is in the direction of a minimal vector in $\mathbf{A}(\mathcal{S}_\infty)$. Sketch $\mathbf{A}(\mathcal{S}_\infty)$ as shown below to see that $\mathbf{v} = (\,1.502 \quad .599\,)^T$ is a maximal vector in $\mathbf{A}(\mathcal{S}_\infty)$.



It's not clear which vector is minimal—don't assume $\begin{pmatrix} .168 & .067 \end{pmatrix}^T$ is. A minimal vector $\mathbf{y}$ in $\mathbf{A}(\mathcal{S}_\infty)$ satisfies $\|\mathbf{y}\|_\infty = \min_{\|\mathbf{x}\|_\infty=1} \|\mathbf{A}\mathbf{x}\|_\infty = 1/\left\|\mathbf{A}^{-1}\right\|_\infty$ (see (5.2.6) on p. 280), so, for $\mathbf{y} = \mathbf{A}\mathbf{x}_0$ with $\|\mathbf{x}_0\|_\infty = 1$,

$$\left\|\mathbf{A}^{-1}\left(\frac{\mathbf{y}}{\|\mathbf{y}\|_\infty}\right)\right\|_\infty = \frac{\|\mathbf{x}_0\|_\infty}{\|\mathbf{y}\|_\infty} = \frac{1}{\|\mathbf{y}\|_\infty} = \left\|\mathbf{A}^{-1}\right\|_\infty = \max_{\|\mathbf{z}\|_\infty=1} \left\|\mathbf{A}^{-1}\mathbf{z}\right\|_\infty .$$

In other words, $\hat{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_\infty$ must be a vector in $\mathcal{S}_\infty$ that receives maximal stretch under $\mathbf{A}^{-1}$. You don't have to look very hard to find such a vector because its components are ± 1—recall the proof of (5.2.15) on p. 283. Notice that $\hat{\mathbf{y}} = \begin{pmatrix} 1 & -1 \end{pmatrix}^T \in \mathcal{S}_\infty$, and $\hat{\mathbf{y}}$ receives maximal stretch under $\mathbf{A}^{-1}$ because $\left\|\mathbf{A}^{-1}\mathbf{y}\right\|_\infty = 1,168,000 = \left\|\mathbf{A}^{-1}\right\|_\infty$, so setting

$$\mathbf{b} = \alpha\mathbf{v} = \alpha\begin{pmatrix} 1.502 \\ .599 \end{pmatrix} \quad \text{and} \quad \mathbf{e} = \beta\hat{\mathbf{y}} = \beta\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

produces equality in (5.12.8), regardless of α and β. You may wish to computationally verify that this is indeed the case.

5.12.10. (a) Consider $\mathbf{A} = \begin{pmatrix} \epsilon & -1 \\ 1 & 0 \end{pmatrix}$ or $\mathbf{A} = \begin{pmatrix} \epsilon & \epsilon^n \\ 0 & \epsilon \end{pmatrix}$ for small $\epsilon \neq 0$.

(b) For $\alpha > 1$, consider

$$\mathbf{A} = \begin{pmatrix} 1 & -\alpha & 0 & \cdots & 0 \\ 0 & 1 & -\alpha & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -\alpha \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{n\times n} \quad \text{and } \mathbf{A}^{-1} = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n-2} & \alpha^{n-1} \\ 0 & 1 & \cdots & \alpha^{n-3} & \alpha^{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \alpha \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Regardless of which norm is used, $\|\mathbf{A}\| > \alpha$ and $\left\|\mathbf{A}^{-1}\right\| > \alpha^{n-1}$, so $\kappa > \alpha^n$ exhibits exponential growth. Even for moderate values of n and $\alpha > 1$, κ can be quite large.

5.12.11. For $\mathbf{B} = \mathbf{A}^{-1}\mathbf{E}$, write $(\mathbf{A} - \mathbf{E}) = \mathbf{A}(\mathbf{I} - \mathbf{B})$, and use the Neumann series expansion to obtain

$$\tilde{\mathbf{x}} = (\mathbf{A}-\mathbf{E})^{-1}\mathbf{b} = (\mathbf{I}-\mathbf{B})^{-1}\mathbf{A}^{-1}\mathbf{b} = (\mathbf{I}+\mathbf{B}+\mathbf{B}^2+\cdots)\mathbf{x} = \mathbf{x}+\mathbf{B}(\mathbf{I}+\mathbf{B}+\mathbf{B}^2+\cdots)\mathbf{x}.$$

Therefore, $\|\mathbf{x} - \tilde{\mathbf{x}}\| \le \|\mathbf{B}\| \sum_{n=0}^{\infty} \|\mathbf{B}\|^n \|\mathbf{x}\| \le \left\|\mathbf{A}^{-1}\right\| \|\mathbf{E}\| \|\mathbf{x}\| \sum_{n=0}^{\infty} \alpha^n$, so

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \left\|\mathbf{A}^{-1}\right\| \|\mathbf{E}\| \frac{1}{1-\alpha} = \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|} \frac{1}{1-\alpha} = \frac{\kappa}{1-\alpha} \frac{\|\mathbf{E}\|}{\|\mathbf{A}\|}.$$

5.12.12. Begin with

$$\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{x} - (\mathbf{I} - \mathbf{B})^{-1}\mathbf{A}^{-1}(\mathbf{b} - \mathbf{e}) = \left(\mathbf{I} - (\mathbf{I} - \mathbf{B})^{-1}\right)\mathbf{x} + (\mathbf{I} - \mathbf{B})^{-1}\mathbf{A}^{-1}\mathbf{e}.$$

Use the triangle inequality with $\mathbf{b} = \mathbf{A}\mathbf{x} \Rightarrow 1/\|\mathbf{x}\| \le \|\mathbf{A}\| / \|\mathbf{b}\|$ to obtain

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \left\|\mathbf{I} - (\mathbf{I} - \mathbf{B})^{-1}\right\| + \left\|(\mathbf{I} - \mathbf{B})^{-1}\right\| \kappa \frac{\|\mathbf{e}\|}{\|\mathbf{b}\|}.$$

Write $(\mathbf{I}-\mathbf{B})^{-1} = \sum_{i=0}^{\infty} \mathbf{B}^i$, and use the identity $\mathbf{I}-(\mathbf{I} - \mathbf{B})^{-1} = -\mathbf{B}(\mathbf{I} - \mathbf{B})^{-1}$ to produce

$$\left\|(\mathbf{I} - \mathbf{B})^{-1}\right\| \le \sum_{i=0}^{\infty} \|\mathbf{B}\|^i = \frac{1}{1 - \|\mathbf{B}\|} \quad \text{and} \quad \left\|\mathbf{I} - (\mathbf{I} - \mathbf{B})^{-1}\right\| \le \frac{\|\mathbf{B}\|}{1 - \|\mathbf{B}\|}.$$

Now combine everything above with $\|\mathbf{B}\| \le \left\|\mathbf{A}^{-1}\right\| \|\mathbf{E}\| = \kappa \|\mathbf{E}\| / \|\mathbf{A}\|$.

5.12.13. Even though the URV factors are not unique, $\mathbf{A}^\dagger$ is, so in each case you should arrive at the same matrix

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{R}^\dagger\mathbf{U}^T = \frac{1}{81}\begin{pmatrix} -4 & 2 & -4 \\ -18 & -18 & 9 \\ -4 & 2 & -4 \\ -2 & 1 & -2 \end{pmatrix}.$$

5.12.14. By (5.12.17), the minimum norm solution is $\mathbf{A}^\dagger\mathbf{b} = (1/9)\begin{pmatrix} 10 & 9 & 10 & 5 \end{pmatrix}^T$.

5.12.15. $\mathbf{U}$ is a unitary matrix in which the columns of $\mathbf{U}_1$ are an orthonormal basis for $R(\mathbf{A})$ and the columns of $\mathbf{U}_2$ are an orthonormal basis for $N\left(\mathbf{A}^T\right)$, so setting $\mathbf{X} = \mathbf{U}_1$, $\mathbf{Y} = \mathbf{U}_2$, and $\left[\mathbf{X} \,|\, \mathbf{Y}\right]^{-1} = \mathbf{U}^T$ in (5.9.12) produces $\mathbf{P} = \mathbf{U}_1\mathbf{U}_1^T$. Furthermore,

$$\mathbf{A}\mathbf{A}^\dagger = \mathbf{U}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T\mathbf{V}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T = \mathbf{U}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T = \mathbf{U}_1\mathbf{U}_1^T.$$

According to (5.9.9), the projector onto $N\left(\mathbf{A}^T\right)$ along $R(\mathbf{A})$ is $\mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{U}_1\mathbf{U}_1^T = \mathbf{U}_2\mathbf{U}_2^T = \mathbf{I} - \mathbf{A}\mathbf{A}^\dagger$.

5.12.16. (a) When $\mathbf{A}$ is nonsingular, $\mathbf{U} = \mathbf{V} = \mathbf{I}$ and $\mathbf{R} = \mathbf{A}$, so $\mathbf{A}^\dagger = \mathbf{A}^{-1}$.

(b) If $\mathbf{A} = \mathbf{URV}^T$ is as given in (5.12.16), where $\mathbf{R} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, it is clear that $(\mathbf{R}^\dagger)^\dagger = \mathbf{R}$, and hence $(\mathbf{A}^\dagger)^\dagger = (\mathbf{VR}^\dagger\mathbf{U}^T)^\dagger = \mathbf{U}(\mathbf{R}^\dagger)^\dagger\mathbf{V}^T = \mathbf{URV}^T = \mathbf{A}$.

(c) For $\mathbf{R}$ as above, it is easy to see that $(\mathbf{R}^\dagger)^T = (\mathbf{R}^T)^\dagger$, so an argument similar to that used in part (b) leads to $(\mathbf{A}^\dagger)^T = (\mathbf{A}^T)^\dagger$.

(d) When $rank\,(\mathbf{A}_{m\times n}) = n$, an SVD must have the form

$$\mathbf{A} = \mathbf{U}_{m\times m}\begin{pmatrix} \mathbf{D}_{n\times n} \\ \mathbf{0}_{m-n\times n} \end{pmatrix}\mathbf{I}_{n\times n}, \quad \text{so} \quad \mathbf{A}^\dagger = \mathbf{I}\,(\,\mathbf{D}^{-1} \quad \mathbf{0}\,)\,\mathbf{U}^T.$$

Furthermore, $\mathbf{A}^T\mathbf{A} = \mathbf{D}^2$, and $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{I}\,(\,\mathbf{D}^{-1} \quad \mathbf{0}\,)\,\mathbf{U}^T = \mathbf{A}^\dagger$. The other part is similar.

(e) $\mathbf{A}^T\mathbf{A}\mathbf{A}^\dagger = \mathbf{V}\begin{pmatrix} \mathbf{C}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T\mathbf{U}\begin{pmatrix} \mathbf{C}_{r\times r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{V}^T\mathbf{V}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T = \mathbf{A}^T$. The other part is similar.

(f) Use an SVD to write

$$\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^\dagger = \mathbf{V}\begin{pmatrix} \mathbf{D}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T\mathbf{U}\begin{pmatrix} \mathbf{D}^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T = \mathbf{V}\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T = \mathbf{A}^\dagger.$$

The other part is similar.

(g) The URV factorization insures that $rank\left(\mathbf{A}^\dagger\right) = rank\,(\mathbf{A}) = rank\left(\mathbf{A}^T\right)$, and part (f) implies $R\left(\mathbf{A}^\dagger\right) \subseteq R\left(\mathbf{A}^T\right)$, so $R\left(\mathbf{A}^\dagger\right) = R\left(\mathbf{A}^T\right)$. Argue that $R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^\dagger\mathbf{A}\right)$ by using Exercise 5.12.15. The other parts are similar.

(h) If $\mathbf{A} = \mathbf{URV}^T$ is a URV factorization for $\mathbf{A}$, then $(\mathbf{PU})\mathbf{R}(\mathbf{Q}^T\mathbf{V})^T$ is a URV factorization for $\mathbf{B} = \mathbf{PAQ}$. So, by (5.12.16), we have

$$\mathbf{B}^\dagger = \mathbf{Q}^T\mathbf{V}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{U}^T\mathbf{P}^T = \mathbf{Q}^T\mathbf{A}^\dagger\mathbf{P}^T.$$

Almost any two singular or rectangular matrices can be used to build a counterexample to show that $(\mathbf{AB})^\dagger$ is not always the same as $\mathbf{B}^\dagger\mathbf{A}^\dagger$.

(i) If $\mathbf{A} = \mathbf{URV}^T$, then $(\mathbf{A}^T\mathbf{A})^\dagger = (\mathbf{VR}^T\mathbf{U}^T\mathbf{URV})^\dagger = \mathbf{V}^T(\mathbf{R}^T\mathbf{R})^\dagger\mathbf{V}^T$. Similarly, $\mathbf{A}^\dagger(\mathbf{A}^T)^\dagger = \mathbf{VR}^\dagger\mathbf{U}^T\mathbf{UR}^{T\dagger}\mathbf{V}^T = \mathbf{VR}^\dagger\mathbf{R}^{T\dagger}\mathbf{V}^T = \mathbf{V}^T(\mathbf{R}^T\mathbf{R})^\dagger\mathbf{V}^T$. The other part is argued in the same way.

5.12.17. If $\mathbf{A}$ is RPN, then $index(\mathbf{A}) = 1$, and the URV decomposition (5.11.15) is a similarity transformation of the kind (5.10.5). That is, $\mathbf{N} = \mathbf{0}$ and $\mathbf{Q} = \mathbf{U}$, so $\mathbf{A}^D$ as defined in (5.10.6) is the same as $\mathbf{A}^\dagger$ as defined by (5.12.16). Conversely, if $\mathbf{A}^\dagger = \mathbf{A}^D$, then

$$\mathbf{A}\mathbf{A}^D = \mathbf{A}^D\mathbf{A} \implies \mathbf{A}^\dagger\mathbf{A} = \mathbf{A}\mathbf{A}^\dagger \implies R\,(\mathbf{A}) = R\left(\mathbf{A}^T\right).$$

5.12.18. (a) Recall that $\|\mathbf{B}\|_F^2 = trace\left(\mathbf{B}^T\mathbf{B}\right)$, and use the fact that $R(\mathbf{X}) \perp R(\mathbf{Y})$ implies $\mathbf{X}^T\mathbf{Y} = \mathbf{0} = \mathbf{Y}^T\mathbf{X}$ to write

$$\begin{aligned}\|\mathbf{X}+\mathbf{Y}\|_F^2 &= trace\left((\mathbf{X}+\mathbf{Y})^T(\mathbf{X}+\mathbf{Y})\right)\\ &= trace\left(\mathbf{X}^T\mathbf{X}+\mathbf{X}^T\mathbf{Y}+\mathbf{Y}^T\mathbf{X}+\mathbf{Y}^T\mathbf{Y}\right)\\ &= trace\left(\mathbf{X}^T\mathbf{X}\right)+trace\left(\mathbf{Y}^T\mathbf{Y}\right) = \|\mathbf{X}\|_F^2+\|\mathbf{Y}\|_F^2 .\end{aligned}$$

(b) Consider $\mathbf{X} = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} 0 & 0 \\ 0 & 3 \end{pmatrix}$.

(c) Use the result of part (a) to write

$$\begin{aligned}\|\mathbf{I}-\mathbf{A}\mathbf{X}\|_F^2 &= \left\|\mathbf{I}-\mathbf{A}\mathbf{A}^\dagger+\mathbf{A}\mathbf{A}^\dagger-\mathbf{A}\mathbf{X}\right\|_F^2\\ &= \left\|\mathbf{I}-\mathbf{A}\mathbf{A}^\dagger\right\|_F^2+\left\|\mathbf{A}\mathbf{A}^\dagger-\mathbf{A}\mathbf{X}\right\|_F^2\\ &\geq \left\|\mathbf{I}-\mathbf{A}\mathbf{A}^\dagger\right\|_F^2 ,\end{aligned}$$

with equality holding if and only if $\mathbf{A}\mathbf{X} = \mathbf{A}\mathbf{A}^\dagger$—i.e., if and only if $\mathbf{X} = \mathbf{A}^\dagger+\mathbf{Z}$, where $R(\mathbf{Z}) \subseteq N(\mathbf{A}) \perp R\left(\mathbf{A}^T\right) = R\left(\mathbf{A}^\dagger\right)$. Moreover, for any such $\mathbf{X}$,

$$\|\mathbf{X}\|_F^2 = \left\|\mathbf{A}^\dagger+\mathbf{Z}\right\|_F^2 = \left\|\mathbf{A}^\dagger\right\|_F^2+\|\mathbf{Z}\|_F^2 \geq \left\|\mathbf{A}^\dagger\right\|_F^2$$

with equality holding if and only if $\mathbf{Z} = \mathbf{0}$.

Solutions for exercises in section 5. 13

5.13.1. $\mathbf{P}_{\mathcal{M}} = \mathbf{u}\mathbf{u}^T/(\mathbf{u}^T\mathbf{u}) = (1/10)\begin{pmatrix} 9 & 3 \\ 3 & 1 \end{pmatrix}$, and $\mathbf{P}_{\mathcal{M}^\perp} = \mathbf{I}-\mathbf{P}_{\mathcal{M}} = (1/10)\begin{pmatrix} 1 & -3 \\ -3 & 9 \end{pmatrix}$, so $\mathbf{P}_{\mathcal{M}}\mathbf{b} = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$, and $\mathbf{P}_{\mathcal{M}^\perp}\mathbf{b} = \begin{pmatrix} -2 \\ 6 \end{pmatrix}$.

5.13.2. (a) Use any of the techniques described in Example 5.13.3 to obtain the following.

$$\mathbf{P}_{R(\mathbf{A})} = \begin{pmatrix} .5 & 0 & .5 \\ 0 & 1 & 0 \\ .5 & 0 & .5 \end{pmatrix} \qquad \mathbf{P}_{N(\mathbf{A})} = \begin{pmatrix} .8 & -.4 & 0 \\ -.4 & .2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{P}_{R(\mathbf{A}^T)} = \begin{pmatrix} .2 & .4 & 0 \\ .4 & .8 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{P}_{N(\mathbf{A}^T)} = \begin{pmatrix} .5 & 0 & -.5 \\ 0 & 0 & 0 \\ -.5 & 0 & .5 \end{pmatrix}$$

(b) The point in $N(\mathbf{A})^\perp$ that is closest to $\mathbf{b}$ is

$$\mathbf{P}_{N(\mathbf{A})^\perp}\mathbf{b} = \mathbf{P}_{R(\mathbf{A}^T)}\mathbf{b} = \begin{pmatrix} .6 \\ 1.2 \\ 1 \end{pmatrix}.$$

5.13.3. If $\mathbf{x} \in R(\mathbf{P})$, then $\mathbf{Px} = \mathbf{x}$—recall (5.9.10)—so $\|\mathbf{Px}\|_2 = \|\mathbf{x}\|_2$. Conversely, suppose $\|\mathbf{Px}\|_2 = \|\mathbf{x}\|_2$, and let $\mathbf{x} = \mathbf{m} + \mathbf{n}$, where $\mathbf{m} \in R(\mathbf{P})$ and $\mathbf{n} \in N(\mathbf{P})$ so that $\mathbf{m} \perp \mathbf{n}$. The Pythagorean theorem (Exercise 5.4.14) guarantees that $\|\mathbf{x}\|_2^2 = \|\mathbf{m}+\mathbf{n}\|_2^2 = \|\mathbf{m}\|_2^2 + \|\mathbf{n}\|_2^2$. But we also have

$$\|\mathbf{x}\|_2^2 = \|\mathbf{Px}\|_2^2 = \|\mathbf{P}(\mathbf{m}+\mathbf{n})\|_2^2 = \|\mathbf{Pm}\|_2^2 = \|\mathbf{m}\|_2^2 .$$

Therefore, $\mathbf{n} = \mathbf{0}$, and thus $\mathbf{x} = \mathbf{m} \in R(\mathbf{P})$.

5.13.4. $(\mathbf{A}^T\mathbf{P}_{R(\mathbf{A})})^T = \mathbf{P}_{R(\mathbf{A})}^T\mathbf{A} = \mathbf{P}_{R(\mathbf{A})}\mathbf{A} = \mathbf{A}$.

5.13.5. Equation (5.13.4) says that $\mathbf{P}_{\mathcal{M}} = \mathbf{U}\mathbf{U}^T = \sum_{i=1}^r \mathbf{u}_i\mathbf{u}_i{}^T$, where $\mathbf{U}$ contains the $\mathbf{u}_i$'s as columns.

5.13.6. The Householder (or Givens) reduction technique can be employed as described in Example 5.11.2 on p. 407 to compute orthogonal matrices $\mathbf{U} = \big(\mathbf{U}_1 \,|\, \mathbf{U}_2\big)$ and $\mathbf{V} = \big(\mathbf{V}_1 \,|\, \mathbf{V}_2\big)$, which are factors in a URV factorization of $\mathbf{A}$. Equation (5.13.12) insures that

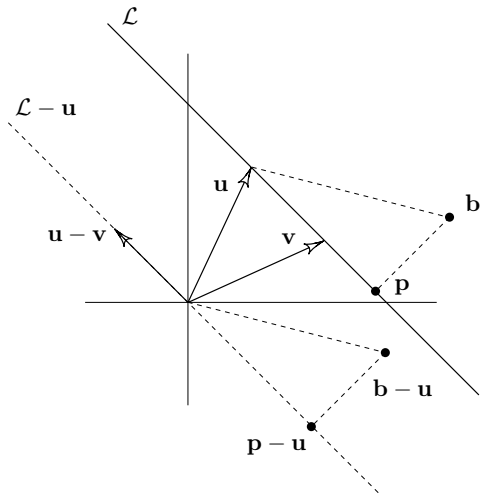
$$\begin{aligned} \mathbf{P}_{R(\mathbf{A})} &= \mathbf{U}_1\mathbf{U}_1^T, & \mathbf{P}_{N(\mathbf{A}^T)} &= \mathbf{P}_{R(\mathbf{A})^\perp} = \mathbf{I} - \mathbf{U}_1\mathbf{U}_1^T = \mathbf{U}_2\mathbf{U}_2^T, \\ \mathbf{P}_{R(\mathbf{A}^T)} &= \mathbf{V}_1\mathbf{V}_1^T, & \mathbf{P}_{N(\mathbf{A})} &= \mathbf{P}_{R(\mathbf{A}^T)^\perp} = \mathbf{I} - \mathbf{V}_1\mathbf{V}_1^T = \mathbf{V}_2\mathbf{V}_2^T. \end{aligned}$$

5.13.7. (a) The only nonsingular orthogonal projector (i.e., the only nonsingular symmetric idempotent matrix) is the identity matrix. Consequently, for all other orthogonal projectors $\mathbf{P}$, we must have $rank\,(\mathbf{P}) = 0$ or $rank\,(\mathbf{P}) = 1$, so $\mathbf{P} = \mathbf{0}$ or, by Example 5.13.1, $\mathbf{P} = (\mathbf{u}\mathbf{u}^T)/\mathbf{u}^T\mathbf{u}$. In other words, the 2×2 orthogonal projectors are $\mathbf{P} = \mathbf{I}$, $\mathbf{P} = \mathbf{0}$, and, for a nonzero vector $\mathbf{u}^T = (\,\alpha \quad \beta\,)$,

$$\mathbf{P} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{\alpha^2+\beta^2}\begin{pmatrix} \alpha^2 & \alpha\beta \\ \alpha\beta & \beta^2 \end{pmatrix}.$$

(b) $\mathbf{P} = \mathbf{I}$, $\mathbf{P} = \mathbf{0}$, and, for nonzero vectors $\mathbf{u}, \mathbf{v} \in \Re^{2\times 1}$, $\mathbf{P} = (\mathbf{u}\mathbf{v}^T)/\mathbf{u}^T\mathbf{v}$.

5.13.8. If either $\mathbf{u}$ or $\mathbf{v}$ is the zero vector, then $\mathcal{L}$ is a one-dimensional subspace, and the solution is given in Example 5.13.1. Suppose that neither $\mathbf{u}$ nor $\mathbf{v}$ is the zero vector, and let $\mathbf{p}$ be the orthogonal projection of $\mathbf{b}$ onto $\mathcal{L}$. Since $\mathcal{L}$ is the translate of the subspace $span\,\{\mathbf{u}-\mathbf{v}\}$, subtracting $\mathbf{u}$ from everything moves the situation back to the origin—the following picture illustrates this in $\Re^2$.



In other words, $\mathcal{L}$ is translated back down to $span\{\mathbf{u}-\mathbf{v}\}$, $\mathbf{b} \to \mathbf{b}-\mathbf{u}$, and $\mathbf{p} \to \mathbf{p}-\mathbf{u}$, so that $\mathbf{p}-\mathbf{u}$ must be the orthogonal projection of $\mathbf{b}-\mathbf{u}$ onto $span\{\mathbf{u}-\mathbf{v}\}$. Example 5.13.1 says that

$$\mathbf{p}-\mathbf{u} = \mathbf{P}_{span\{\mathbf{u}-\mathbf{v}\}}(\mathbf{b}-\mathbf{u}) = \frac{(\mathbf{u}-\mathbf{v})(\mathbf{u}-\mathbf{v})^T}{(\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})}(\mathbf{b}-\mathbf{u}),$$

and thus

$$\mathbf{p} = \mathbf{u} + \left[\frac{(\mathbf{u}-\mathbf{v})^T(\mathbf{b}-\mathbf{u})}{(\mathbf{u}-\mathbf{v})^T(\mathbf{u}-\mathbf{v})}\right](\mathbf{u}-\mathbf{v}).$$

5.13.9. $\|\mathbf{A}\widehat{\mathbf{x}}-\mathbf{b}\|_2 = \left\|\mathbf{P}_{R(\mathbf{A})}\mathbf{b}-\mathbf{b}\right\|_2 = \left\|(\mathbf{I}-\mathbf{P}_{R(\mathbf{A})})\mathbf{b}\right\|_2 = \left\|\mathbf{P}_{N(\mathbf{A}^T)}\mathbf{b}\right\|_2$

5.13.10. Use (5.13.17) with $\mathbf{P}_{R(\mathbf{A})} = \mathbf{P}_{R(\mathbf{A})}^T = \mathbf{P}_{R(\mathbf{A})}^2$, to write

$$\begin{aligned}\|\boldsymbol{\varepsilon}\|_2^2 &= (\mathbf{b}-\mathbf{P}_{R(\mathbf{A})}\mathbf{b})^T(\mathbf{b}-\mathbf{P}_{R(\mathbf{A})}\mathbf{b})\\ &= \mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{P}_{R(\mathbf{A})}^T\mathbf{b} - \mathbf{b}^T\mathbf{P}_{R(\mathbf{A})}\mathbf{b} + \mathbf{b}^T\mathbf{P}_{R(\mathbf{A})}^T\mathbf{P}_{R(\mathbf{A})}\mathbf{b}\\ &= \mathbf{b}^T\mathbf{b} - \mathbf{b}^T\mathbf{P}_{R(\mathbf{A})}\mathbf{b} = \|\mathbf{b}\|_2^2 - \left\|\mathbf{P}_{R(\mathbf{A})}\mathbf{b}\right\|_2^2.\end{aligned}$$

5.13.11. According to (5.13.13) we must show that $\sum_{i=1}^r(\mathbf{u}_i{}^T\mathbf{x})\mathbf{u}_i = \mathbf{P}_{\mathcal{M}}\mathbf{x}$. It follows from (5.13.4) that if $\mathbf{U}_{n\times r}$ is the matrix containing the vectors in $\mathcal{B}$ as columns, then

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}\mathbf{U}^T = \sum_{i=1}^r \mathbf{u}_i\mathbf{u}_i{}^T \quad\Longrightarrow\quad \mathbf{P}_{\mathcal{M}}\mathbf{x} = \sum_{i=1}^r \mathbf{u}_i\mathbf{u}_i{}^T\mathbf{x} = \sum_{i=1}^r(\mathbf{u}_i{}^T\mathbf{x})\mathbf{u}_i.$$

5.13.12. Yes, the given spanning set $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an orthonormal basis for $\mathcal{M}$, so, by Exercise 5.13.11,

$$\mathbf{P}_{\mathcal{M}}\mathbf{b} = \sum_{i=1}^3(\mathbf{u}_i{}^T\mathbf{b})\mathbf{u}_i = \mathbf{u}_1 + 3\mathbf{u}_2 + 7\mathbf{u}_3 = \begin{pmatrix}5\\0\\5\\3\end{pmatrix}.$$

5.13.13. (a) Combine the fact that $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $R(\mathbf{P}_{\mathcal{N}}) \subseteq N(\mathbf{P}_{\mathcal{M}})$ with the facts $R(\mathbf{P}_{\mathcal{N}}) = \mathcal{N}$ and $N(\mathbf{P}_{\mathcal{M}}) = \mathcal{M}^{\perp}$ to write

$$\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0} \Longleftrightarrow \mathcal{N} \subseteq \mathcal{M}^{\perp} \Longleftrightarrow \mathcal{N} \perp \mathcal{M}.$$

(b) Yes—this is a direct consequence of part (a). Alternately, you could say

$$\mathbf{0} = \mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} \Longleftrightarrow \mathbf{0} = (\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}})^T = \mathbf{P}_{\mathcal{N}}^T\mathbf{P}_{\mathcal{M}}^T = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}.$$

5.13.14. (a) Use Exercise 4.2.9 along with (4.5.5) to write

$$\begin{aligned} R(\mathbf{P}_{\mathcal{M}}) + R(\mathbf{P}_{\mathcal{N}}) &= R(\mathbf{P}_{\mathcal{M}} \mid \mathbf{P}_{\mathcal{N}}) = R\left((\mathbf{P}_{\mathcal{M}} \mid \mathbf{P}_{\mathcal{N}}) \begin{pmatrix} \mathbf{P}_{\mathcal{M}} \\ \hline \mathbf{P}_{\mathcal{N}} \end{pmatrix}^T \right) \\ &= R\left(\mathbf{P}_{\mathcal{M}}{\mathbf{P}_{\mathcal{M}}}^T + \mathbf{P}_{\mathcal{N}}{\mathbf{P}_{\mathcal{N}}}^T \right) = R\left(\mathbf{P}_{\mathcal{M}}^2 + \mathbf{P}_{\mathcal{N}}^2\right) \\ &= R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}). \end{aligned}$$

(b) $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0} \Longleftrightarrow R(\mathbf{P}_{\mathcal{N}}) \subseteq N(\mathbf{P}_{\mathcal{M}}) \Longleftrightarrow \mathcal{N} \subseteq \mathcal{M}^{\perp} \Longleftrightarrow \mathcal{M} \perp \mathcal{N}$.

(c) Exercise 5.9.17 says $\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}$ is idempotent if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}}$. Because $\mathbf{P}_{\mathcal{M}}$ and $\mathbf{P}_{\mathcal{N}}$ are symmetric, $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{0}$ if and only if $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{0}$ (via the reverse order law for transposition). The fact that $R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = R(\mathbf{P}_{\mathcal{M}}) \oplus R(\mathbf{P}_{\mathcal{N}}) = \mathcal{M} \oplus \mathcal{N}$ was established in Exercise 5.9.17, and $\mathcal{M} \perp \mathcal{N}$ follows from part (b).

5.13.15. First notice that $\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}$ is symmetric, so (5.13.12) and the result of Exercise 5.13.14, part (a), can be combined to conclude that

$$(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger} = (\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}) = \mathbf{P}_{R(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})} = \mathbf{P}_{\mathcal{M}+\mathcal{N}}.$$

Now, $\mathcal{M} \subseteq \mathcal{M} + \mathcal{N}$ implies $\mathbf{P}_{\mathcal{M}+\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}}$, and the reverse order law for transposition yields $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{M}+\mathcal{N}} = \mathbf{P}_{\mathcal{M}}$ so that $\mathbf{P}_{\mathcal{M}+\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{M}+\mathcal{N}}$. In other words, $(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})$, or

$$\begin{aligned} &\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}} \\ &\qquad = \mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}}. \end{aligned}$$

Subtracting $\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$ from both sides of this equation produces

$$\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}} + \mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}.$$

Let $\mathbf{Z} = 2\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}} = 2\mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}$, and notice that $R(\mathbf{Z}) \subseteq R(\mathbf{P}_{\mathcal{M}}) = \mathcal{M}$ and $R(\mathbf{Z}) \subseteq R(\mathbf{P}_{\mathcal{N}}) = \mathcal{N}$ implies $R(\mathbf{Z}) \subseteq \mathcal{M} \cap \mathcal{N}$. Furthermore, $\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = \mathbf{P}_{\mathcal{N}}\mathbf{P}_{\mathcal{M}\cap\mathcal{N}}$, and $\mathbf{P}_{\mathcal{M}+\mathcal{N}}\mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}$, so, by the

reverse order law for transposition, $\mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{P}_{\mathcal{N}}$ and $\mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{P}_{\mathcal{M}+\mathcal{N}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}$. Consequently,

$$\begin{aligned}\mathbf{Z} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{Z} &= \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\left[\mathbf{P}_{\mathcal{M}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{N}} + \mathbf{P}_{\mathcal{N}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}\mathbf{P}_{\mathcal{M}}\right]\\ &= \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}})^{\dagger}(\mathbf{P}_{\mathcal{M}}+\mathbf{P}_{\mathcal{N}}) = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}\mathbf{P}_{\mathcal{M}+\mathcal{N}} = \mathbf{P}_{\mathcal{M}\cap\mathcal{N}}.\end{aligned}$$

5.13.16. (a) Use the fact that $\mathbf{A}^T = \mathbf{A}^T\mathbf{P}_{R(\mathbf{A})} = \mathbf{A}^T\mathbf{A}\mathbf{A}^{\dagger}$ (see Exercise 5.13.4) to write

$$\begin{aligned}\int_0^{\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\mathbf{A}^T dt &= \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\mathbf{A}^T\mathbf{A}\mathbf{A}^{\dagger} dt = \left(\int_0^{\infty} \mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\mathbf{A}^T\mathbf{A} dt\right)\mathbf{A}^{\dagger}\\ &= \Big[-\mathrm{e}^{-\mathbf{A}^T\mathbf{A}t}\Big]_0^{\infty}\mathbf{A}^{\dagger} = [\mathbf{0}-(-\mathbf{I})]\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}.\end{aligned}$$

(b) Recall from Example 5.10.5 that $\mathbf{A}^k = \mathbf{A}^{k+1}\mathbf{A}^D = \mathbf{A}^k\mathbf{A}\mathbf{A}^D$, and write

$$\begin{aligned}\int_0^{\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t}\mathbf{A}^k dt &= \int_0^{\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t}\mathbf{A}^k\mathbf{A}\mathbf{A}^D dt = \left(\int_0^{\infty} \mathrm{e}^{-\mathbf{A}^{k+1}t}\mathbf{A}^{k+1}\mathbf{A} dt\right)\mathbf{A}^D\\ &= \Big[-\mathrm{e}^{-\mathbf{A}^{k+1}t}\Big]_0^{\infty}\mathbf{A}^D = [\mathbf{0}-(-\mathbf{I})]\mathbf{A}^D = \mathbf{A}^D.\end{aligned}$$

(c) This is just a special case of the formula in part (b) with $k = 0$. However, it is easy to derive the formula directly by writing

$$\begin{aligned}\int_0^{\infty} \mathrm{e}^{-\mathbf{A}t} dt &= \int_0^{\infty} \mathrm{e}^{-\mathbf{A}t}\mathbf{A}\mathbf{A}^{-1} dt = \left(\int_0^{\infty} \mathrm{e}^{-\mathbf{A}t}\mathbf{A} dt\right)\mathbf{A}^{-1}\\ &= \Big[\mathrm{e}^{-\mathbf{A}t}\Big]_0^{\infty}\mathbf{A}^{-1} = [\mathbf{0}-(-\mathbf{I})]\mathbf{A}^{-1} = \mathbf{A}^{-1}.\end{aligned}$$

5.13.17. (a) The points in $\mathcal{H}$ are just solutions to a linear system $\mathbf{u}^T\mathbf{x} = \beta$. Using the fact that the general solution of any linear system is a particular solution plus the general solution of the associated homogeneous equation produces

$$\mathcal{H} = \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}} + N(\mathbf{u}^T) = \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}} + [R(\mathbf{u})]^{\perp} = \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}} + \mathbf{u}^{\perp},$$

where $\mathbf{u}^{\perp}$ denotes the orthogonal complement of the one-dimensional space spanned by the vector $\mathbf{u}$. Thus $\mathcal{H} = \mathbf{v}+\mathcal{M}$, where $\mathbf{v} = \beta\mathbf{u}/\mathbf{u}^T\mathbf{u}$ and $\mathcal{M} = \mathbf{u}^{\perp}$. The fact that $\dim(\mathbf{u}^{\perp}) = n-1$ follows directly from (5.11.3).

(b) Use (5.13.14) with part (a) and the fact that $\mathbf{P}_{\mathbf{u}^{\perp}} = \mathbf{I} - \mathbf{u}\mathbf{u}^T/\mathbf{u}^T\mathbf{u}$ to write

$$\mathbf{p} = \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}} + \left(\mathbf{I} - \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\right)\left(\mathbf{b} - \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}}\right) = \frac{\beta\mathbf{u}}{\mathbf{u}^T\mathbf{u}} + \mathbf{b} - \frac{\mathbf{u}\mathbf{u}^T\mathbf{b}}{\mathbf{u}^T\mathbf{u}} = \mathbf{b} - \left(\frac{\mathbf{u}^T\mathbf{b}-\beta}{\mathbf{u}^T\mathbf{u}}\right)\mathbf{u}.$$

5.13.18. (a) $\mathbf{u}^T\mathbf{w} \neq 0$ implies $\mathcal{M} \cap \mathcal{W} = \mathbf{0}$ so that

$$\dim(\mathcal{M} + \mathcal{W}) = \dim \mathcal{M} + \dim \mathcal{W} = (n-1) + 1 = n.$$

Therefore, $\mathcal{M} + \mathcal{W} = \Re^n$. This together with $\mathcal{M} \cap \mathcal{W} = \mathbf{0}$ means $\Re^n = \mathcal{M} \oplus \mathcal{W}$.

(b) Write

$$\mathbf{b} = \mathbf{b} - \frac{\mathbf{u}^T\mathbf{b}}{\mathbf{u}^T\mathbf{w}}\mathbf{w} + \frac{\mathbf{u}^T\mathbf{b}}{\mathbf{u}^T\mathbf{w}}\mathbf{w} = \mathbf{p} + \frac{\mathbf{u}^T\mathbf{b}}{\mathbf{u}^T\mathbf{w}}\mathbf{w},$$

and observe that $\mathbf{p} \in \mathcal{M}$ (because $\mathbf{u}^T\mathbf{p} = 0$) and $(\mathbf{u}^T\mathbf{b}/\mathbf{u}^T\mathbf{w})\mathbf{w} \in \mathcal{W}$. By definition, $\mathbf{p}$ is the projection of $\mathbf{b}$ onto $\mathcal{M}$ along $\mathcal{W}$.

(c) We know from Exercise 5.13.17, part (a), that $\mathcal{H} = \mathbf{v} + \mathcal{M}$, where $\mathbf{v} = \beta\mathbf{u}/\mathbf{u}^T\mathbf{u}$ and $\mathcal{M} = \mathbf{u}^\perp$, so subtracting $\mathbf{v} = \beta\mathbf{u}/\mathbf{u}^T\mathbf{u}$ from everything in $\mathcal{H}$ as well as from $\mathbf{b}$ translates the situation back to the origin. Sketch a picture similar to that of Figure 5.13.5 to see that this moves $\mathcal{H}$ back to $\mathcal{M}$, it translates $\mathbf{b}$ to $\mathbf{b}-\mathbf{v}$, and it translates $\mathbf{p}$ to $\mathbf{p}-\mathbf{v}$. Now, $\mathbf{p}-\mathbf{v}$ should be the projection of $\mathbf{b}-\mathbf{v}$ onto $\mathcal{M}$ along $\mathcal{W}$, so by the result of part (b),

$$\mathbf{p}-\mathbf{v} = \mathbf{b}-\mathbf{v}-\frac{\mathbf{u}^T(\mathbf{b}-\mathbf{v})}{\mathbf{u}^T\mathbf{w}}\mathbf{w} \implies \mathbf{p} = \mathbf{b}-\frac{\mathbf{u}^T(\mathbf{b}-\mathbf{v})}{\mathbf{u}^T\mathbf{w}}\mathbf{w} = \mathbf{b}-\left(\frac{\mathbf{u}^T\mathbf{b}-\beta}{\mathbf{u}^T\mathbf{w}}\right)\mathbf{w}.$$

5.13.19. For convenience, set $\beta = \mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i$ so that $\mathbf{p}_{kn+i} = \mathbf{p}_{kn+i-1} - \beta(\mathbf{A}_{i*})^T$. Use the fact that

$$\mathbf{A}_{i*}(\mathbf{p}_{kn+i-1} - \mathbf{x}) = \mathbf{A}_{i*}\mathbf{p}_{kn+i-1} - b_i = \beta$$

together with $\|\mathbf{A}_{i*}\|_2 = 1$ to write

$$\begin{aligned}
\|\mathbf{p}_{kn+i} - \mathbf{x}\|_2^2 &= \left\|\mathbf{p}_{kn+i-1} - \beta(\mathbf{A}_{i*})^T - \mathbf{x}\right\|_2^2 \\
&= \left\|(\mathbf{p}_{kn+i-1} - \mathbf{x}) - \beta(\mathbf{A}_{i*})^T\right\|_2^2 \\
&= (\mathbf{p}_{kn+i-1} - \mathbf{x})^T(\mathbf{p}_{kn+i-1} - \mathbf{x}) \\
&\qquad - 2\beta\mathbf{A}_{i*}(\mathbf{p}_{kn+i-1} - \mathbf{x}) + \beta^2\mathbf{A}_{i*}(\mathbf{A}_{i*})^T \\
&= \|\mathbf{p}_{kn+i-1} - \mathbf{x}\|_2^2 - \beta^2.
\end{aligned}$$

Consequently, $\|\mathbf{p}_{kn+i} - \mathbf{x}\|_2 \leq \|\mathbf{p}_{kn+i-1} - \mathbf{x}\|_2$, with equality holding if and only if $\beta = 0$ or, equivalently, if and only if $\mathbf{p}_{kn+i-1} \in \mathcal{H}_{i-1} \cap \mathcal{H}_i$. Therefore, the sequence of norms $\|\mathbf{p}_{kn+i} - \mathbf{x}\|_2$ is monotonically decreasing, and hence it must have a limiting value. This implies that the sequence of the β's defined above must approach 0, and thus the sequence of the $\mathbf{p}_{kn+i}$'s converges to $\mathbf{x}$.

5.13.20. Refer to Figure 5.13.8, and notice that the line passing from $\mathbf{p}_1^{(1)}$ to $\mathbf{p}_2^{(1)}$ is parallel to $\mathcal{V} = span\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right)$, so projecting $\mathbf{p}_1^{(1)}$ through $\mathbf{p}_2^{(1)}$ onto $\mathcal{H}_2$

is exactly the same as projecting $\mathbf{p}_1^{(1)}$ onto $\mathcal{H}_2$ along (i.e., parallel to) $\mathcal{V}$. According to part (c) of Exercise 5.13.18, this projection is given by

$$\begin{aligned}\mathbf{p}_2^{(2)} &= \mathbf{p}_1^{(1)} - \frac{\mathbf{A}_{2*}\left(\mathbf{p}_1^{(1)} - b_1\mathbf{A}_{2*}^T\right)}{\mathbf{A}_{2*}\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right)}\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right)\\ &= \mathbf{p}_1^{(1)} - \frac{\left(\mathbf{A}_{2*}\mathbf{p}_1^{(1)} - b_1\right)}{\mathbf{A}_{2*}\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right)}\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right).\end{aligned}$$

All other projections are similarly derived. It is now straightforward to verify that the points created by the algorithm are exactly the same points described in Steps $1, 2, \ldots, n-1$.

Note: The condition that $\left\{\left(\mathbf{p}_1^{(1)} - \mathbf{p}_2^{(1)}\right), \left(\mathbf{p}_1^{(1)} - \mathbf{p}_3^{(1)}\right), \ldots, \left(\mathbf{p}_1^{(1)} - \mathbf{p}_n^{(1)}\right)\right\}$ is independent insures that $\left\{\left(\mathbf{p}_2^{(2)} - \mathbf{p}_3^{(2)}\right), \left(\mathbf{p}_2^{(2)} - \mathbf{p}_4^{(2)}\right), \ldots, \left(\mathbf{p}_2^{(2)} - \mathbf{p}_n^{(2)}\right)\right\}$ is also independent. The same holds at each subsequent step. Furthermore, $\mathbf{A}_{2*}\left(\mathbf{p}_1^{(1)} - \mathbf{p}_k^{(1)}\right) \neq 0$ for $k > 1$ implies that $\mathcal{V}_k = span\left(\mathbf{p}_1^{(1)} - \mathbf{p}_k^{(1)}\right)$ is not parallel to $\mathcal{H}_2$, so all projections onto $\mathcal{H}_2$ along $\mathcal{V}_k$ are well defined. It can be argued that the analogous situation holds at each step of the process—i.e., the initial conditions insure $\mathbf{A}_{i+1*}\left(\mathbf{p}_i^{(i)} - \mathbf{p}_k^{(i)}\right) \neq 0$ for $k > i$.

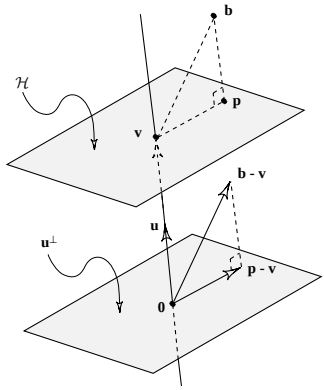
5.13.21. Equation (5.13.13) says that the orthogonal distance between $\mathbf{x}$ and $\mathcal{M}^\perp$ is

$$dist\left(\mathbf{x}, \mathcal{M}^\perp\right) = \left\|\mathbf{x} - \mathbf{P}_{\mathcal{M}^\perp}\mathbf{x}\right\|_2 = \left\|(\mathbf{I} - \mathbf{P}_{\mathcal{M}^\perp})\mathbf{x}\right\|_2 = \left\|\mathbf{P}_{\mathcal{M}}\mathbf{x}\right\|_2.$$

Similarly,

$$dist\left(\mathbf{R}\mathbf{x}, \mathcal{M}^\perp\right) = \left\|\mathbf{P}_{\mathcal{M}}\mathbf{R}\mathbf{x}\right\|_2 = \left\|-\mathbf{P}_{\mathcal{M}}\mathbf{x}\right\|_2 = \left\|\mathbf{P}_{\mathcal{M}}\mathbf{x}\right\|_2.$$

5.13.22. (a) We know from Exercise 5.13.17 that $\mathcal{H} = \mathbf{v} + \mathbf{u}^\perp$, where $\mathbf{v} = \beta\mathbf{u}$, so subtracting $\mathbf{v}$ from everything in $\mathcal{H}$ as well as from $\mathbf{b}$ translates the situation back to the origin. As depicted in the diagram below, this moves $\mathcal{H}$ down to $\mathbf{u}^\perp$, and it translates $\mathbf{b}$ to $\mathbf{b} - \mathbf{v}$ and $\mathbf{r}$ to $\mathbf{r} - \mathbf{v}$.



Now, we know from (5.6.8) that the reflection of $\mathbf{b}-\mathbf{v}$ about $\mathbf{u}^\perp$ is

$$\mathbf{r}-\mathbf{v}=\mathbf{R}(\mathbf{b}-\mathbf{v})=(\mathbf{I}-2\mathbf{u}\mathbf{u}^T)(\mathbf{b}-\mathbf{v})=\mathbf{b}+(\beta-2\mathbf{u}^T\mathbf{b})\mathbf{u},$$

and therefore the reflection of $\mathbf{b}$ about $\mathcal{H}$ is

$$\mathbf{r}=\mathbf{R}(\mathbf{b}-\mathbf{v})+\mathbf{v}=\mathbf{b}-2(\mathbf{u}^T\mathbf{b}-\beta)\mathbf{u}.$$

(b) From part (a), the reflection of $\mathbf{r}_0$ about $\mathcal{H}_i$ is

$$\mathbf{r}_i=\mathbf{r}_0-2(\mathbf{A}_{i*}\mathbf{r}_0-b_i)\left(\mathbf{A}_{i*}\right)^T,$$

and therefore the mean value of all of the reflections $\{\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n\}$ is

$$\begin{aligned}\mathbf{m}&=\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{r}_0-2(\mathbf{A}_{i*}\mathbf{r}_0-b_i)\left(\mathbf{A}_{i*}\right)^T\right)\\&=\mathbf{r}_0-\frac{2}{n}\sum_{i=1}^{n}(\mathbf{A}_{i*}\mathbf{r}_0-b_i)(\mathbf{A}_{i*})^T\\&=\mathbf{r}_0-\frac{2}{n}\mathbf{A}^T\left(\mathbf{A}\mathbf{r}_0-\mathbf{b}\right)=\mathbf{r}_0-\frac{2}{n}\mathbf{A}^T\boldsymbol{\varepsilon}.\end{aligned}$$

Note: If weights $w_i>0$ such that $\sum w_i=1$ are used, then the weighted mean is

$$\begin{aligned}\mathbf{m}&=\sum_{i=1}^{n}w_i\mathbf{r}_i=\sum_{i=1}^{n}w_i\left(\mathbf{r}_0-2(\mathbf{A}_{i*}\mathbf{r}_0-b_i)\left(\mathbf{A}_{i*}\right)^T\right)\\&=\mathbf{r}_0-2\sum_{i=1}^{n}w_i(\mathbf{A}_{i*}\mathbf{r}_0-b_i)(\mathbf{A}_{i*})^T\\&=\mathbf{r}_0-\frac{2}{n}\mathbf{A}^T\mathbf{W}\left(\mathbf{A}\mathbf{r}_0-\mathbf{b}\right)=\mathbf{r}_0-\frac{2}{n}\mathbf{A}^T\mathbf{W}\boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{W} = diag\{w_1, w_2, \ldots, w_n\}$.

(c) First observe that

$$\begin{aligned}
\mathbf{x} - \mathbf{m}_k &= \mathbf{x} - \mathbf{m}_{k-1} + \frac{2}{n}\mathbf{A}^T\boldsymbol{\varepsilon}_{k-1} \\
&= \mathbf{x} - \mathbf{m}_{k-1} + \frac{2}{n}\mathbf{A}^T(\mathbf{A}\mathbf{m}_{k-1} - \mathbf{b}) \\
&= \mathbf{x} - \mathbf{m}_{k-1} + \frac{2}{n}\mathbf{A}^T(\mathbf{A}\mathbf{m}_{k-1} - \mathbf{A}\mathbf{x}) \\
&= \mathbf{x} - \mathbf{m}_{k-1} + \frac{2}{n}\mathbf{A}^T\mathbf{A}(\mathbf{m}_{k-1} - \mathbf{x}) \\
&= \left(\mathbf{I} - \frac{2}{n}\mathbf{A}^T\mathbf{A}\right)(\mathbf{x} - \mathbf{m}_{k-1}),
\end{aligned}$$

and then use successive substitution to conclude that

$$\mathbf{x} - \mathbf{m}_k = \left(\mathbf{I} - \frac{2}{n}\mathbf{A}^T\mathbf{A}\right)^k (\mathbf{x} - \mathbf{m}_0).$$

Solutions for exercises in section 5. 14

5.14.1. Use (5.14.5) to observe that

$$E[y_i y_j] = \mathrm{Cov}[y_i, y_j] + \mu_{y_i}\mu_{y_j} = \begin{cases} \sigma^2 + (\mathbf{X}_{i*}\boldsymbol{\beta})^2 & \text{if } i = j, \\ (\mathbf{X}_{i*}\boldsymbol{\beta})(\mathbf{X}_{j*}\boldsymbol{\beta}) & \text{if } i \neq j, \end{cases}$$

so that

$$E[\mathbf{y}\mathbf{y}^T] = \sigma^2\mathbf{I} + (\mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta})^T = \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T.$$

Write $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y}$, and use the fact that $\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger$ is idempotent to obtain

$$\hat{\mathbf{e}}^T\hat{\mathbf{e}} = \mathbf{y}^T(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y} = trace\left((\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y}\mathbf{y}^T\right).$$

Now use the linearity of trace and expectation together with the result of Exercise 5.9.13 and the fact that $(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{X} = \mathbf{0}$ to write

$$\begin{aligned}
E[\hat{\mathbf{e}}^T\hat{\mathbf{e}}] &= E\left[trace\left((\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y}\mathbf{y}^T\right)\right] = trace\left(E[(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)\mathbf{y}\mathbf{y}^T]\right) \\
&= trace\left((\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)E[\mathbf{y}\mathbf{y}^T]\right) = trace\left((\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger)(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T)\right) \\
&= \sigma^2 trace\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\dagger\right) = \sigma^2\left(m - trace\left(\mathbf{X}\mathbf{X}^\dagger\right)\right) \\
&= \sigma^2\left(m - rank\left(\mathbf{X}\mathbf{X}^\dagger\right)\right) = \sigma^2(m - n).
\end{aligned}$$

Solutions for exercises in section 5. 15

5.15.1. (a) $\theta_{min} = 0$, and $\theta_{max} = \theta = \phi = \pi/4$.

(b) $\theta_{min} = \theta = \phi = \pi/4$, and $\theta_{max} = 1$.

5.15.2. (a) The first principal angle is $\theta_1 = \theta_{min} = 0$, and we can take $\mathbf{u}_1 = \mathbf{v}_1 = \mathbf{e}_1$. This means that

$$\mathcal{M}_2 = \mathbf{u}_1^{\perp} \cap \mathcal{M} = span\{\mathbf{e}_2\} \quad \text{and} \quad \mathcal{N}_2 = \mathbf{v}_1^{\perp} \cap \mathcal{N} = span\{(0,\,1,\,1)\}.$$

The second principal angle is the minimal angle between $\mathcal{M}_2$ and $\mathcal{N}_2$, and this is just the angle between $\mathbf{e}_2$ and $(0,\,1,\,1)$, so $\theta_2 = \pi/4$.

(b) This time the first principal angle is $\theta_1 = \theta_{min} = \pi/4$, and we can take $\mathbf{u}_1 = \mathbf{e}_1$ and $\mathbf{v}_1 = (0,\,1/\sqrt{2},\,1/\sqrt{2})$. There are no more principal angles because $\mathcal{N}_2 = \mathbf{v}_1^{\perp} \cap \mathcal{N} = \mathbf{0}$.

5.15.3. (a) This follows from (5.15.16) because $\mathbf{P}_{\mathcal{M}} = \mathbf{P}_{\mathcal{N}}$ if and only if $\mathcal{M} = \mathcal{N}$.

(b) If $\mathbf{0} \neq \mathbf{x} \in \mathcal{M} \cap \mathcal{N}$, then (5.15.1) evaluates to 1 with the maximum being attained at $\mathbf{u} = \mathbf{v} = \mathbf{x}/\|\mathbf{x}\|_2$. Conversely, $\cos\theta_{min} = 1 \implies \mathbf{v}^T\mathbf{u} = 1$ for some $\mathbf{u} \in \mathcal{M}$ and $\mathbf{v} \in \mathcal{N}$ such that $\|\mathbf{u}\|_2 = 1 = \|\mathbf{v}\|_2$. But $\mathbf{v}^T\mathbf{u} = 1 = \|\mathbf{u}\|_2\,\|\mathbf{v}\|_2$ represents equality in the CBS inequality (5.1.3), and we know this occurs if and only if $\mathbf{v} = \alpha\mathbf{u}$ for $\alpha = \mathbf{v}^T\mathbf{u}/\mathbf{u}^*\mathbf{u} = 1/1 = 1$. Thus $\mathbf{u} = \mathbf{v} \in \mathcal{M} \cap \mathcal{N}$.

(c) $\max\limits_{\substack{\mathbf{u}\in\mathcal{M},\,\mathbf{v}\in\mathcal{N} \\ \|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}} \mathbf{v}^T\mathbf{u} = 0 \iff \mathbf{v}^T\mathbf{u} = 0\ \forall\ \mathbf{u} \in \mathcal{M},\ \mathbf{v} \in \mathcal{V} \iff \mathcal{M} \perp \mathcal{N}$.

5.15.4. You can use either (5.15.3) or (5.15.4) to arrive at the result. The latter is used by observing

$$\begin{aligned}\left\|(\mathbf{P}_{\mathcal{M}^\perp} - \mathbf{P}_{\mathcal{N}^\perp})^{-1}\right\|_2 &= \left\|\left((\mathbf{I} - \mathbf{P}_{\mathcal{M}}) - (\mathbf{I} - \mathbf{P}_{\mathcal{N}})\right)^{-1}\right\|_2 \\ &= \left\|(\mathbf{P}_{\mathcal{N}} - \mathbf{P}_{\mathcal{M}})^{-1}\right\|_2 = \left\|(\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}})^{-1}\right\|_2.\end{aligned}$$

5.15.5. $\mathcal{M} \oplus \mathcal{N}^\perp = \Re^n \implies \dim\mathcal{M} = \dim\mathcal{N} \implies \sin\theta_{max} = \delta(\mathcal{M},\mathcal{N}) = \delta(\mathcal{N},\mathcal{M})$, so $\cos\tilde{\theta}_{min} = \|\mathbf{P}_{\mathcal{M}}\mathbf{P}_{\mathcal{N}^\perp}\|_2 = \|\mathbf{P}_{\mathcal{M}}(\mathbf{I} - \mathbf{P}_{\mathcal{N}})\|_2 = \delta(\mathcal{M},\mathcal{N}) = \sin\theta_{max}$.

5.15.6. It was argued in the proof of (5.15.4) that $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular whenever $\mathcal{M}$ and $\mathcal{N}$ are complementary, so we need only prove the converse. Suppose $\dim\mathcal{M} = r > 0$ and $\dim\mathcal{N} = k > 0$ (the problem is trivial if $r = 0$ or $k = 0$) so that $\mathbf{U}_1^T\mathbf{V}_1$ is $r \times n-k$ and $\mathbf{U}_2^T\mathbf{V}_2$ is $n-r \times k$. If $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{N}}$ is nonsingular, then (5.15.7) insures that the rows as well as the columns in each of these products must be linearly independent. That is, $\mathbf{U}_1^T\mathbf{V}_1$ and $\mathbf{U}_2^T\mathbf{V}_2$ must both be square and nonsingular, so $r + k = n$. Combine this with the formula for the rank of a product (4.5.1) to conclude

$$\begin{aligned}k &= rank\left(\mathbf{U}_2^T\mathbf{V}_2\right) = rank\left(\mathbf{U}_2^T\right) - \dim N\left(\mathbf{U}_2^T\right) \cap R\left(\mathbf{V}_2\right) \\ &= n - r - \dim\mathcal{M} \cap \mathcal{N} = k - \dim\mathcal{M} \cap \mathcal{N}.\end{aligned}$$

It follows that $\mathcal{M} \cap \mathcal{N} = \mathbf{0}$, and hence $\mathcal{M} \oplus \mathcal{N} = \Re^n$.

5.15.7. (a) This can be derived from (5.15.7), or it can be verified by direct multiplication by using $\mathbf{P}_{\mathcal{N}}(\mathbf{I}-\mathbf{P}) = \mathbf{I}-\mathbf{P} \implies \mathbf{P}-\mathbf{P}_{\mathcal{N}}\mathbf{P} = \mathbf{I}-\mathbf{P}_{\mathcal{N}}$ to write

$$\begin{aligned}(\mathbf{P}_{\mathcal{M}}-\mathbf{P}_{\mathcal{N}})(\mathbf{P}-\mathbf{Q}) &= \mathbf{P}_{\mathcal{M}}\mathbf{P}-\mathbf{P}_{\mathcal{M}}\mathbf{Q}-\mathbf{P}_{\mathcal{N}}\mathbf{P}+\mathbf{P}_{\mathcal{N}}\mathbf{Q}\\ &= \mathbf{P}-\mathbf{0}-\mathbf{P}_{\mathcal{N}}\mathbf{P}+\mathbf{P}_{\mathcal{N}}\mathbf{Q} = \mathbf{I}-\mathbf{P}_{\mathcal{N}}+\mathbf{P}_{\mathcal{N}}\mathbf{Q}\\ &= \mathbf{I}-\mathbf{P}_{\mathcal{N}}(\mathbf{I}-\mathbf{Q}) = \mathbf{I}.\end{aligned}$$

(b) and (c) follow from (a) in conjunction with (5.15.3) and (5.15.4).

5.15.8. Since we are maximizing over a larger set, $\max_{\|\mathbf{x}\|=1} f(\mathbf{x}) \le \max_{\|\mathbf{x}\|\le 1} f(\mathbf{x})$. A strict inequality here implies the existence of a nonzero vector $\mathbf{x}_0$ such that $\|\mathbf{x}_0\| < 1$ and $f(\mathbf{x}) < f(\mathbf{x}_0)$ for all vectors such that $\|\mathbf{x}\| = 1$. But then

$$f(\mathbf{x}_0) > f(\mathbf{x}_0/\|\mathbf{x}_0\|) = f(\mathbf{x}_0)/\|\mathbf{x}_0\| \implies \|\mathbf{x}_0\|\, f(\mathbf{x}_0) > f(\mathbf{x}_0),$$

which is impossible because $\|\mathbf{x}_0\| < 1$.

5.15.9. (a) We know from equation (5.15.6) that $\mathbf{P}_{\mathcal{MN}} = \mathbf{U}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^T$ in which $\mathbf{C}$ is nonsingular and $\mathbf{C}^{-1} = \mathbf{V}_1^T\mathbf{U}_1$. Consequently,

$$\mathbf{P}_{\mathcal{MN}}^{\dagger} = \mathbf{V}\begin{pmatrix}\mathbf{C}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{U}^T = \mathbf{V}_1\mathbf{C}^{-1}\mathbf{U}_1^T = \mathbf{V}_1\mathbf{V}_1^T\mathbf{U}_1\mathbf{U}_1^T = \mathbf{P}_{\mathcal{N}^{\perp}}\mathbf{P}_{\mathcal{M}}.$$

(b) Use the fact

$$\begin{aligned}\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 &= \left\|(\mathbf{V}_1^T\mathbf{U}_1)^{-1}\right\|_2 = \left\|\mathbf{U}_1(\mathbf{V}_1^T\mathbf{U}_1)^{-1}\mathbf{V}_1^T\right\|_2\\ &= \left\|(\mathbf{V}_1\mathbf{V}_1^T\mathbf{U}_1\mathbf{U}_1^T)^{\dagger}\right\|_2 = \left\|\left[(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\right]^{\dagger}\right\|_2\end{aligned}$$

(and similarly for the other term) to show that

$$\left\|\left[(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\right]^{\dagger}\right\|_2 = \left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|\left[\mathbf{P}_{\mathcal{M}}(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\right]^{\dagger}\right\|_2,$$

and

$$\left\|\left[(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\mathbf{P}_{\mathcal{N}}\right]^{\dagger}\right\|_2 = \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2 = \left\|\left[\mathbf{P}_{\mathcal{N}}(\mathbf{I}-\mathbf{P}_{\mathcal{M}})\right]^{\dagger}\right\|_2.$$

It was established in the proof of (5.15.4) that $\left\|(\mathbf{U}_1^T\mathbf{V}_1)^{-1}\right\|_2 = \left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2$, so combining this with the result of part (a) and (5.15.3) produces the desired conclusion.

5.15.10. (a) We know from (5.15.2) that $\cos\bar{\theta}_{min} = \|\mathbf{P}_{\mathcal{N}^{\perp}}\mathbf{P}_{\mathcal{M}}\|_2 = \|(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\|_2$, and we know from Exercise 5.15.9 that $\mathbf{P}_{\mathcal{MN}} = \left[(\mathbf{I}-\mathbf{P}_{\mathcal{N}})\mathbf{P}_{\mathcal{M}}\right]^{\dagger}$, so taking the pseudoinverse of both sides of this yields the desired result.

(b) Use (5.15.3) together with part (a), (5.13.10), and (5.13.12) to write

$$1 = \left\|\mathbf{P}_{\mathcal{MN}}\mathbf{P}^{\dagger}_{\mathcal{MN}}\right\|_2 \leq \left\|\mathbf{P}_{\mathcal{MN}}\right\|_2 \left\|\mathbf{P}^{\dagger}_{\mathcal{MN}}\right\|_2 = \frac{\cos\bar{\theta}_{min}}{\sin\theta_{min}}.$$

5.15.11. (a) Use the facts that $\|\mathbf{A}\|_2 = \|\mathbf{A}^T\|_2$ and $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$ to write

$$\begin{aligned}
\frac{1}{\left\|(\mathbf{U}_2^T\mathbf{V}_2)^{-1}\right\|_2^2} &= \frac{1}{\left\|(\mathbf{V}_2^T\mathbf{U}_2)^{-1}\right\|_2^2} = \min_{\|\mathbf{x}\|_2=1} \left\|\mathbf{V}_2^T\mathbf{U}_2\mathbf{x}\right\|_2^2 \\
&= \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{U}_2^T\mathbf{V}_2\mathbf{V}_2^T\mathbf{U}_2\mathbf{x} \\
&= \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{U}_2^T(\mathbf{I}-\mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_2\mathbf{x} = \min_{\|\mathbf{x}\|_2=1} \left(1 - \left\|\mathbf{V}_1^T\mathbf{U}_2\mathbf{x}\right\|_2^2\right) \\
&= 1 - \max_{\|\mathbf{x}\|_2=1} \left\|\mathbf{V}_1^T\mathbf{U}_2\mathbf{x}\right\|_2^2 = 1 - \left\|\mathbf{V}_1^T\mathbf{U}_2\right\|_2^2 = 1 - \left\|\mathbf{U}_2^T\mathbf{V}_1\right\|_2^2.
\end{aligned}$$

(b) Use a similar technique to write

$$\begin{aligned}
\left\|\mathbf{U}_2^T\mathbf{V}_2\right\|_2^2 &= \left\|\mathbf{U}_2^T\mathbf{V}_2\mathbf{V}_2^T\right\|_2^2 = \left\|\mathbf{U}_2^T(\mathbf{I}-\mathbf{V}_1\mathbf{V}_1^T)\right\|_2^2 \\
&= \left\|(\mathbf{I}-\mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_2\right\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{U}_2^T(\mathbf{I}-\mathbf{V}_1\mathbf{V}_1^T)\mathbf{U}_2\mathbf{x} \\
&= 1 - \min_{\|\mathbf{x}\|_2=1} \left\|\mathbf{V}_1^T\mathbf{U}_2\mathbf{x}\right\|_2^2 = 1 - \frac{1}{\left\|(\mathbf{V}_1^T\mathbf{U}_2)^{-1}\right\|_2^2} \\
&= 1 - \frac{1}{\left\|(\mathbf{U}_2^T\mathbf{V}_1)^{-1}\right\|_2^2}.
\end{aligned}$$

Solutions for Chapter 6

Solutions for exercises in section 6. 1

6.1.1. (a) -1 (b) 8 (c) $-\alpha\beta\gamma$

(d) $a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - (a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33} + a_{13}a_{22}a_{31})$

(This is where the "diagonal rule" you learned in high school comes from.)

6.1.2. If $\mathbf{A} = [\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3]$, then $V_3 = \left[\det\left(\mathbf{A}^T\mathbf{A}\right)\right]^{1/2} = 20$ (recall Example 6.1.4). But you could also realize that the $\mathbf{x}_i$'s are mutually orthogonal to conclude that $V_3 = \|\mathbf{x}_1\|_2 \|\mathbf{x}_2\|_2 \|\mathbf{x}_3\|_2 = 20$.

6.1.3. (a) 10 (b) 0 (c) 120 (d) 39 (e) 1 (f) $(n-1)!$

6.1.4. $rank\,(\mathbf{A}) = 2$

6.1.5. A square system has a unique solution if and only if its coefficient matrix is nonsingular—recall the discussion in §2.5. Consequently, (6.1.13) guarantees that a square system has a unique solution if and only if the determinant of the coefficient matrix is nonzero. Since

$$\begin{vmatrix} 1 & \alpha & 0 \\ 0 & 1 & -1 \\ \alpha & 0 & 1 \end{vmatrix} = 1 - \alpha^2,$$

it follows that there is a unique solution if and only if $\alpha \neq \pm 1$.

6.1.6. $\mathbf{I} = \mathbf{A}^{-1}\mathbf{A} \implies \det(\mathbf{I}) = \det\left(\mathbf{A}^{-1}\mathbf{A}\right) = \det\left(\mathbf{A}^{-1}\right)\det(\mathbf{A})$

$\implies 1 = \det\left(\mathbf{A}^{-1}\right)\det(\mathbf{A}) \implies \det\left(\mathbf{A}^{-1}\right) = 1/\det(\mathbf{A})$.

6.1.7. Use the product rule (6.1.15) to write

$$\begin{aligned}\det\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right) &= \det\left(\mathbf{P}^{-1}\right)\det(\mathbf{A})\det(\mathbf{P}) = \det\left(\mathbf{P}^{-1}\right)\det(\mathbf{P})\det(\mathbf{A}) \\ &= \det\left(\mathbf{P}^{-1}\mathbf{P}\right)\det(\mathbf{A}) = \det(\mathbf{I})\det(\mathbf{A}) = \det(\mathbf{A}).\end{aligned}$$

6.1.8. Use (6.1.4) together with the fact that $\overline{z_1 z_2} = \bar{z}_1\bar{z}_2$ and $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$ for all complex numbers to write

$$\begin{aligned}\det(\mathbf{A}^*) = \det\left(\bar{\mathbf{A}}^T\right) = \det\left(\bar{\mathbf{A}}\right) &= \sum_p \sigma(p)\overline{a_{1p_1}}\cdots\overline{a_{np_n}} \\ &= \sum_p \sigma(p)\overline{a_{1p_1}\cdots a_{np_n}} = \overline{\sum_p \sigma(p) a_{1p_1}\cdots a_{np_n}} = \overline{\det(\mathbf{A})}.\end{aligned}$$

6.1.9. (a) $\mathbf{I} = \mathbf{Q}^*\mathbf{Q} \implies 1 = \det(\mathbf{Q}^*\mathbf{Q}) = \det(\mathbf{Q}^*)\det(\mathbf{Q}) = [\det(\mathbf{Q})]^2$ by Exercise 6.1.8.

(b) If $\mathbf{A} = \mathbf{UDV}^*$ is an SVD, then, by part (a),

$$\begin{aligned}|\det(\mathbf{A})| &= |\det(\mathbf{UDV}^*)| = |\det(\mathbf{U})|\,|\det(\mathbf{D})|\,|\det(\mathbf{V}^*)| \\ &= \det(\mathbf{D}) = \sigma_1\sigma_2\cdots\sigma_n.\end{aligned}$$

6.1.10. Let $r = rank(\mathbf{A})$, and let $\sigma_1 \geq \cdots \geq \sigma_r$ be the nonzero singular values of $\mathbf{A}$. If $\mathbf{A} = \mathbf{U}_{m\times m}\begin{pmatrix}\mathbf{D}_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}_{m\times n}(\mathbf{V}^*)_{n\times n}$ is an SVD, then, by Exercises 6.1.9 and 6.1.8, $\det(\mathbf{V})\det(\mathbf{V}^*) = |\det(\mathbf{V})|^2 = 1$, so

$$\begin{aligned}\det(\mathbf{A}^*\mathbf{A}) = \det(\mathbf{VD}^*\mathbf{DV}^*) &= \det(\mathbf{V})\begin{vmatrix}(\mathbf{D}^*\mathbf{D})_{r\times r} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{vmatrix}_{n\times n}\det(\mathbf{V}^*) \\ &= \sigma_1^2\sigma_2^2\cdots\sigma_r^2\,\underbrace{0\cdots0}_{n-r}, \quad \text{and this is} \quad \begin{cases}=0 & \text{when } r<n,\\ >0 & \text{when } r=n.\end{cases}\end{aligned}$$

Note: You can't say $\det(\mathbf{A}^*\mathbf{A}) = \overline{\det(\mathbf{A})}\det(\mathbf{A}) = |\det(\mathbf{A})|^2 \geq 0$ because $\mathbf{A}$ need not be square.

6.1.11. $\alpha\mathbf{A} = (\alpha\mathbf{I})\mathbf{A} \implies \det(\alpha\mathbf{A}) = \det(\alpha\mathbf{I})\det(\mathbf{A}) = \alpha^n\det(\mathbf{A})$.

6.1.12. $\mathbf{A} = -\mathbf{A}^T \implies \det(\mathbf{A}) = \det\left(-\mathbf{A}^T\right) = \det(-\mathbf{A}) = (-1)^n\det(\mathbf{A})$ (by Exercise 6.1.11) $\implies \det(\mathbf{A}) = -\det(\mathbf{A})$ when n is odd $\implies \det(\mathbf{A}) = 0$.

6.1.13. If $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is lower triangular and $\mathbf{U}$ is upper triangular where each has 1's on its diagonal and random integers in the remaining nonzero positions, then $\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) = 1\times 1 = 1$, and the entries of $\mathbf{A}$ are rather random integers.

6.1.14. According to the definition,

$$\begin{aligned}\det(\mathbf{A}) &= \sum_p \sigma(p)a_{1p_1}\cdots a_{kp_k}\cdots a_{np_n} \\ &= \sum_p \sigma(p)a_{1p_1}\cdots(x_{p_k}+y_{p_k}+\cdots+z_{p_k})\cdots a_{np_n} \\ &= \sum_p \sigma(p)a_{1p_1}\cdots x_{p_k}\cdots a_{np_n} + \sum_p \sigma(p)a_{1p_1}\cdots y_{p_k}\cdots a_{np_n} \\ &\quad + \cdots + \sum_p \sigma(p)a_{1p_1}\cdots z_{p_k}\cdots a_{np_n} \\ &= \det\begin{pmatrix}\mathbf{A}_{1*}\\ \vdots\\ \mathbf{x}^T\\ \vdots\\ \mathbf{A}_{n*}\end{pmatrix} + \det\begin{pmatrix}\mathbf{A}_{1*}\\ \vdots\\ \mathbf{y}^T\\ \vdots\\ \mathbf{A}_{n*}\end{pmatrix} + \cdots + \det\begin{pmatrix}\mathbf{A}_{1*}\\ \vdots\\ \mathbf{z}^T\\ \vdots\\ \mathbf{A}_{n*}\end{pmatrix}.\end{aligned}$$

6.1.15. If $\mathbf{A}_{n\times 2} = [\mathbf{x}\,|\,\mathbf{y}]$, then the result of Exercise 6.1.10 implies

$$\begin{aligned} 0 \le \det(\mathbf{A}^*\mathbf{A}) &= \begin{vmatrix} \mathbf{x}^*\mathbf{x} & \mathbf{x}^*\mathbf{y} \\ \mathbf{y}^*\mathbf{x} & \mathbf{y}^*\mathbf{y} \end{vmatrix} = (\mathbf{x}^*\mathbf{x})(\mathbf{y}^*\mathbf{y}) - (\mathbf{x}^*\mathbf{y})(\mathbf{y}^*\mathbf{x}) \\ &= \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - (\mathbf{x}^*\mathbf{y})\,\overline{(\mathbf{x}^*\mathbf{y})} \\ &= \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2 - |\mathbf{x}^*\mathbf{y}|^2, \end{aligned}$$

with equality holding if and only if $rank\,(\mathbf{A}) < 2$—i.e., if and only if $\mathbf{y}$ is a scalar multiple of $\mathbf{x}$.

6.1.16. Partition $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{LU} = \begin{pmatrix} \mathbf{L}_k & \mathbf{0} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{U}_k & \mathbf{U}_{12} \\ \mathbf{0} & \mathbf{U}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_k\mathbf{U}_k & * \\ * & * \end{pmatrix}$$

to deduce that $\mathbf{A}_k$ can be written in the form

$$\mathbf{A}_k = \mathbf{L}_k\mathbf{U}_k = \begin{pmatrix} \mathbf{L}_{k-1} & \mathbf{0} \\ \mathbf{d}^T & 1 \end{pmatrix} \begin{pmatrix} \mathbf{U}_{k-1} & \mathbf{c} \\ \mathbf{0} & u_{kk} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_{k-1} = \mathbf{L}_{k-1}\mathbf{U}_{k-1}.$$

The product rule (6.1.15) shows that

$$\det(\mathbf{A}_k) = \det(\mathbf{U}_{k-1}) \times u_{kk} = \det(\mathbf{A}_{k-1}) \times u_{kk},$$

and the desired conclusion follows.

6.1.17. According to (3.10.12), a matrix has an LU factorization if and only if each leading principal submatrix is nonsingular. The leading $k \times k$ principal submatrix of $\mathbf{A}^T\mathbf{A}$ is given by $\mathbf{P}_k = \mathbf{A}_k^T\mathbf{A}_k$, where $\mathbf{A}_k = [\mathbf{A}_{*1}\,|\,\mathbf{A}_{*2}\,|\,\cdots\,|\,\mathbf{A}_{*k}]$. If $\mathbf{A}$ has full column rank, then any nonempty subset of columns is linearly independent, so $rank\,(\mathbf{A}_k) = k$. Therefore, the results of Exercise 6.1.10 insure that $\det(\mathbf{P}_k) = \det\left(\mathbf{A}_k^T\mathbf{A}_k\right) > 0$ for each k, and hence $\mathbf{A}^T\mathbf{A}$ has an LU factorization. The fact that each pivot is positive follows from Exercise 6.1.16.

6.1.18. (a) To evaluate $\det(\mathbf{A})$, use Gaussian elimination as shown below.

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

$$\longrightarrow \begin{pmatrix} 1 & -1 & 3-x \\ 0 & 4-x & -5 \\ 0 & 5-x & -x^2+5x-2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 & 3-x \\ 0 & 4-x & -5 \\ 0 & 0 & \frac{x^3-9x^2+17x+17}{4-x} \end{pmatrix} = \mathbf{U}.$$

Since one interchange was used, $\det(\mathbf{A})$ is (-1) times the product of the diagonal entries of $\mathbf{U}$, so

$$\det(\mathbf{A}) = -x^3 + 9x^2 - 17x - 17 \quad \text{and} \quad \frac{d\Big(\det(\mathbf{A})\Big)}{dx} = -3x^2 + 18x - 17.$$

(b) Using formula (6.1.19) produces

$$\frac{d\Big(\det(\mathbf{A})\Big)}{dx} = \begin{vmatrix} -1 & 0 & 0 \\ 0 & 4-x & -5 \\ 1 & -1 & 3-x \end{vmatrix} + \begin{vmatrix} 2-x & 3 & 4 \\ 0 & -1 & 0 \\ 1 & -1 & 3-x \end{vmatrix} + \begin{vmatrix} 2-x & 3 & 4 \\ 0 & 4-x & -5 \\ 0 & 0 & -1 \end{vmatrix}$$

$$= (-x^2+7x-7) + (-x^2+5x-2) + (-x^2+6x-8)$$
$$= -3x^2+18x-17.$$

6.1.19. No—almost any 2×2 example will show that this cannot hold in general.

6.1.20. It was argued in Example 4.3.6 that if there is at least one value of x for which the Wronski matrix

$$\mathbf{W}(x) = \begin{pmatrix} f_1(x) & f_2(x) & \cdots & f_n(x) \\ f_1'(x) & f_2'(x) & \cdots & f_n'(x) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^{(n-1)}(x) & f_2^{(n-1)}(x) & \cdots & f_n^{(n-1)}(x) \end{pmatrix}$$

is nonsingular, then $\mathcal{S}$ is a linearly independent set. This is equivalent to saying that if $\mathcal{S}$ is a linearly *dependent* set, then the Wronski matrix $\mathbf{W}(x)$ is singular for all values of x. But (6.1.14) insures that a matrix is singular if and only if its determinant is zero, so, if $\mathcal{S}$ is linearly dependent, then the Wronskian $w(x)$ must vanish for every value of x. The converse of this statement is false (Exercise 4.3.14).

6.1.21. (a) $(n!)(n-1)$ (b) 11×11 (c) About 9.24×10^{153} sec $\approx 3 \times 10^{146}$ years (d) About 3×10^{150} mult/sec. (Now this would truly be a "super computer.")

Solutions for exercises in section 6. 2

6.2.1. (a) 8 (b) 39 (c) -3

6.2.2. (a) $\mathbf{A}^{-1} = \dfrac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})} = \dfrac{1}{8}\begin{pmatrix} 0 & 1 & -1 \\ -8 & 4 & 4 \\ 16 & -6 & -2 \end{pmatrix}$

(b) $\mathbf{A}^{-1} = \dfrac{\operatorname{adj}(\mathbf{A})}{\det(\mathbf{A})} = \dfrac{1}{39}\begin{pmatrix} -12 & 25 & -14 & 7 \\ -9 & 9 & 9 & 15 \\ -6 & 6 & 6 & -3 \\ 9 & 4 & 4 & -2 \end{pmatrix}$

6.2.3. (a) $x_1 = 1-\beta, \quad x_2 = \alpha+\beta-1, \quad x_3 = 1-\alpha$

(b) Cramer's rule yields

$$x_2(t) = \frac{\begin{vmatrix} 1 & t^4 & t^2 \\ t^2 & t^3 & t \\ t & 0 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & t & t^2 \\ t^2 & 1 & t \\ t & t^2 & 1 \end{vmatrix}} = \frac{t\begin{vmatrix} t^4 & t^2 \\ t^3 & t \end{vmatrix} + \begin{vmatrix} 1 & t^4 \\ t^2 & t^3 \end{vmatrix}}{\begin{vmatrix} 1 & t \\ t^2 & 1 \end{vmatrix} - t\begin{vmatrix} t^2 & t \\ t & 1 \end{vmatrix} + t^2\begin{vmatrix} t^2 & 1 \\ t & t^2 \end{vmatrix}}$$

$$= \frac{t^3 - t^6}{(t^3-1)(t^3-1)} = \frac{-t^3}{(t^3-1)},$$

and hence

$$\lim_{t\to\infty} x_2(t) = \lim_{t\to\infty} \frac{-1}{1-1/t^3} = -1.$$

6.2.4. Yes.

6.2.5. (a) Almost any two matrices will do the job. One example is $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = -\mathbf{I}$.

(b) Again, almost anything you write down will serve the purpose. One example is $\mathbf{A} = \mathbf{D} = \mathbf{0}_{2\times 2}, \mathbf{B} = \mathbf{C} = \mathbf{I}_{2\times 2}$.

6.2.6. Recall from Example 5.13.3 that $\mathbf{Q} = \mathbf{I} - \mathbf{B}\mathbf{B}^T\mathbf{B}^{-1}\mathbf{B}^T$. According to (6.2.1),

$$\det\left(\mathbf{A}^T\mathbf{A}\right) = \det\begin{pmatrix} \mathbf{B}^T\mathbf{B} & \mathbf{B}^T\mathbf{c} \\ \mathbf{c}^T\mathbf{B} & \mathbf{c}^T\mathbf{c} \end{pmatrix} = \det\left(\mathbf{B}^T\mathbf{B}\right)\left(\mathbf{c}^T\mathbf{Q}\mathbf{c}\right).$$

Since $\det\left(\mathbf{B}^T\mathbf{B}\right) > 0$ (by Exercise 6.1.10), $\mathbf{c}^T\mathbf{Q}\mathbf{c} = \det\left(\mathbf{A}^T\mathbf{A}\right)/\det\left(\mathbf{B}^T\mathbf{B}\right)$.

6.2.7. Expand $\begin{vmatrix} \mathbf{A} & -\mathbf{C} \\ \mathbf{D}^T & \mathbf{I}_k \end{vmatrix}$ both of the ways indicated in (6.2.1).

6.2.8. The result follows from Example 6.2.8, which says $\mathbf{A}[\operatorname{adj}(\mathbf{A})] = \det(\mathbf{A})\,\mathbf{I}$, together with the fact that $\mathbf{A}$ is singular if and only if $\det(\mathbf{A}) = 0$.

6.2.9. The solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, and Example 6.2.7 says that the entries in $\mathbf{A}^{-1}$ are continuous functions of the entries in $\mathbf{A}$. Since $x_i = \sum_k [\mathbf{A}^{-1}]_{ik} b_k$, and since the sum of continuous functions is again continuous, it follows that each x_i is a continuous function of the a_{ij}'s.

6.2.10. If $\mathbf{B} = \alpha\mathbf{A}$, then Exercise 6.1.11 implies $\mathring{B}_{ij} = \alpha^{n-1}\mathring{A}_{ij}$, so $\mathring{\mathbf{B}} = \alpha^{n-1}\mathring{\mathbf{A}}$, and hence $\operatorname{adj}(\mathbf{B}) = \alpha^{n-1}\operatorname{adj}(\mathbf{A})$.

6.2.11. (a) We saw in §6.1 that $rank(\mathbf{A})$ is the order of the largest nonzero minor of $\mathbf{A}$. If $rank(\mathbf{A}) < n-1$, then every minor of order $n-1$ (as well as $\det(\mathbf{A})$ itself) must be zero. Consequently, $\mathring{\mathbf{A}} = \mathbf{0}$, and thus $\operatorname{adj}(\mathbf{A}) = \mathring{\mathbf{A}}^T = \mathbf{0}$.

(b)
$$\begin{aligned} rank(\mathbf{A}) = n-1 \quad &\Longrightarrow \quad \text{at least one minor of order } n-1 \text{ is nonzero} \\ &\Longrightarrow \quad \text{some } \mathring{A}_{ij} \neq 0 \quad \Longrightarrow \quad \operatorname{adj}(\mathbf{A}) \neq \mathbf{0} \\ &\Longrightarrow \quad rank(\operatorname{adj}(\mathbf{A})) \geq 1. \end{aligned}$$

$$\begin{aligned}
\text{Also,}\quad rank\,(\mathbf{A}) = n-1 \quad &\Longrightarrow \quad \det(\mathbf{A}) = 0 \\
&\Longrightarrow \quad \mathbf{A}[\operatorname{adj}(\mathbf{A})] = \mathbf{0} \qquad \text{(by Exercise 6.2.8)} \\
&\Longrightarrow \quad R(\operatorname{adj}(\mathbf{A})) \subseteq N(\mathbf{A}) \\
&\Longrightarrow \quad \dim R(\operatorname{adj}(\mathbf{A})) \leq \dim N(\mathbf{A}) \\
&\Longrightarrow \quad rank\,(\operatorname{adj}(\mathbf{A})) \leq n - rank\,(\mathbf{A}) = 1.
\end{aligned}$$

(c) $rank\,(\mathbf{A}) = n \implies \det(\mathbf{A}) \neq 0 \implies \operatorname{adj}(\mathbf{A}) = \det(\mathbf{A})\,\mathbf{A}^{-1}$
$\implies rank\,(\operatorname{adj}(\mathbf{A})) = n$

6.2.12. If $\det(\mathbf{A}) = 0$, then Exercise 6.2.11 insures that $rank\,(\operatorname{adj}(\mathbf{A})) \leq 1$. Consequently, $\det(\operatorname{adj}(\mathbf{A})) = 0$, and the result is trivially true because both sides are zero. If $\det(\mathbf{A}) \neq 0$, apply the product rule (6.1.15) to $\mathbf{A}[\operatorname{adj}(\mathbf{A})] = \det(\mathbf{A})\,\mathbf{I}$ (from Example 6.2.8) to obtain $\det(\mathbf{A})\det(\operatorname{adj}(\mathbf{A})) = [\det(\mathbf{A})]^n$, so that $\det(\operatorname{adj}(\mathbf{A})) = [\det(\mathbf{A})]^{n-1}$.

6.2.13. Expanding in terms of cofactors of the first row produces $D_n = 2\mathring{A}_{11} - \mathring{A}_{12}$. But $\mathring{A}_{11} = D_{n-1}$ and expansion using the first column yields

$$\mathring{A}_{12} = (-1)\begin{vmatrix} -1 & -1 & 0 & \cdots & 0 \\ 0 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{vmatrix} = (-1)(-1)D_{n-2},$$

so $D_n = 2D_{n-1} - D_{n-2}$. By recursion (or by direct substitution), it is easy to see that the solution of this equation is $D_n = n+1$.

6.2.14. (a) Use the results of Example 6.2.1 with $\lambda_i = 1/\alpha_i$.

(b) Recognize that the matrix $\mathbf{A}$ is a rank-one updated matrix in the sense that

$$\mathbf{A} = (\alpha - \beta)\mathbf{I} + \beta\mathbf{e}\mathbf{e}^T, \quad \text{where} \quad \mathbf{e} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

If $\alpha = \beta$, then $\mathbf{A}$ is singular, so $\det(\mathbf{A}) = 0$. If $\alpha \neq \beta$, then (6.2.3) may be applied to obtain

$$\det(\mathbf{A}) = \det\Big((\alpha-\beta)\mathbf{I}\Big)\left(1 + \frac{\beta\mathbf{e}^T\mathbf{e}}{\alpha-\beta}\right) = (\alpha-\beta)^n\left(1 + \frac{n\beta}{\alpha-\beta}\right).$$

(c) Recognize that the matrix is $\mathbf{I} + \mathbf{e}\mathbf{d}^T$, where

$$\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{d} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

Apply (6.2.2) to produce the desired formula.

6.2.15. (a) Use the second formula in (6.2.1).

(b) Apply the first formula in (6.2.1) along with (6.2.7).

6.2.16. If $\lambda = 0$, then the result is trivially true because both sides are zero. If $\lambda \neq 0$, then expand $\begin{vmatrix} \lambda \mathbf{I}_m & \lambda \mathbf{B} \\ \mathbf{C} & \lambda \mathbf{I}_n \end{vmatrix}$ both of the ways indicated in (6.2.1).

6.2.17. (a) Use the product rule (6.1.15) together with (6.2.2) to write

$$\mathbf{A} + \mathbf{c}\mathbf{d}^T = \mathbf{A} + \mathbf{A}\mathbf{x}\mathbf{d}^T = \mathbf{A}\left(\mathbf{I} + \mathbf{x}\mathbf{d}^T\right).$$

(b) Apply the same technique used in part (a) to obtain

$$\mathbf{A} + \mathbf{c}\mathbf{d}^T = \mathbf{A} + \mathbf{c}\mathbf{y}^T\mathbf{A} = \left(\mathbf{I} + \mathbf{c}\mathbf{y}^T\right)\mathbf{A}.$$

6.2.18. For an elementary reflector $\mathbf{R} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^T/\mathbf{u}^T\mathbf{u}$, (6.2.2) insures $\det(\mathbf{R}) = -1$. If $\mathbf{A}_{n\times n}$ is reduced to upper-triangular form (say $\mathbf{PA} = \mathbf{T}$) by Householder reduction as explained on p. 341, then $\det(\mathbf{P})\det(\mathbf{A}) = \det(\mathbf{T}) = t_{11}\cdots t_{nn}$. Since $\mathbf{P}$ is the product of elementary reflectors, $\det(\mathbf{A}) = (-1)^k t_{11}\cdots t_{nn}$, where k is the number of reflections used in the reduction process. In general, one reflection is required to annihilate entries below a diagonal position, so, if no reduction steps can be skipped, then $\det(\mathbf{A}) = (-1)^{n-1}t_{11}\cdots t_{nn}$. If $\mathbf{P}_{ij}$ is a plane rotation, then there is a permutation matrix (a product of interchange matrices) $\mathbf{B}$ such that $\mathbf{P}_{ij} = \mathbf{B}^T\begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\mathbf{B}$, where $\mathbf{Q} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ with $c^2 + s^2 = 1$. Consequently, $\det(\mathbf{P}_{ij}) = \det\left(\mathbf{B}^T\right)\begin{vmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{vmatrix}\det(\mathbf{B}) = \det(\mathbf{Q}) = 1$ because $\det(\mathbf{B})\det\left(\mathbf{B}^T\right) = \det(\mathbf{B})^2 = 1$ by (6.1.9). Since Givens reduction produces $\mathbf{PA} = \mathbf{T}$, where $\mathbf{P}$ is a product of plane rotations and $\mathbf{T}$ is upper triangular, the product rule (6.1.15) insures $\det(\mathbf{P}) = 1$, so $\det(\mathbf{A}) = \det(\mathbf{T}) = t_{11}\cdots t_{nn}$.

6.2.19. If $\det(\mathbf{A}) = \pm 1$, then (6.2.7) implies $\mathbf{A}^{-1} = \pm\operatorname{adj}(\mathbf{A})$, and thus $\mathbf{A}^{-1}$ is an integer matrix because the cofactors are integers. Conversely, if $\mathbf{A}^{-1}$ is an integer matrix, then $\det\left(\mathbf{A}^{-1}\right)$ and $\det(\mathbf{A})$ are both integers. Since

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} \implies \det(\mathbf{A})\det\left(\mathbf{A}^{-1}\right) = 1,$$

it follows that $\det(\mathbf{A}) = \pm 1$.

6.2.20. (a) Exercise 6.2.19 guarantees that $\mathbf{A}^{-1}$ has integer entries if and only if $\det(\mathbf{A}) = \pm 1$, and (6.2.2) says that $\det(\mathbf{A}) = 1 - 2\mathbf{v}^T\mathbf{u}$, so $\mathbf{A}^{-1}$ has integer entries if and only if $\mathbf{v}^T\mathbf{u}$ is either 0 or 1.

(b) According to (3.9.1),

$$\mathbf{A}^{-1} = \left(\mathbf{I} - 2\mathbf{u}\mathbf{v}^T\right)^{-1} = \mathbf{I} - \frac{2\mathbf{u}\mathbf{v}^T}{2\mathbf{v}^T\mathbf{u} - 1},$$

and thus $\mathbf{A}^{-1} = \mathbf{A}$ when $\mathbf{v}^T\mathbf{u} = 1$.

6.2.21. For $n = 2$, two multiplications are required, and $c(2) = 2$. Assume $c(k)$ multiplications are required to evaluate any $k \times k$ determinant by cofactors. For a $k+1 \times k+1$ matrix, the cofactor expansion in terms of the ith row is

$$\det(\mathbf{A}) = a_{i1}\mathring{A}_{i1} + \cdots + a_{ik}\mathring{A}_{ik} + a_{ik+1}\mathring{A}_{ik+1}.$$

Each $\mathring{A}_{ij}$ requires $c(k)$ multiplications, so the above expansion contains

$$\begin{aligned}(k+1) + (k+1)c(k) &= (k+1) + (k+1)k!\left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(k-1)!}\right)\\ &= (k+1)!\left(\frac{1}{k!} + \left(1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{(k-1)!}\right)\right)\\ &= c(k+1)\end{aligned}$$

multiplications. Remember that $e^x = 1 + x + x^2/2! + x^3/3! + \cdots$, so for $n = 100$,

$$1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{99!} \approx e - 1,$$

and $c(100) \approx 100!(e-1)$. Consequently, approximately 1.6×10^{152} seconds (i.e., 5.1×10^{144} years) are required.

6.2.22. $\mathbf{A} - \lambda\mathbf{I}$ is singular if and only if $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. The cofactor expansion in terms of the first row yields

$$\begin{aligned}\det(\mathbf{A} - \lambda\mathbf{I}) &= -\lambda\begin{vmatrix} 5-\lambda & 2 \\ -3 & -\lambda \end{vmatrix} + 3\begin{vmatrix} 2 & 2 \\ -2 & -\lambda \end{vmatrix} - 2\begin{vmatrix} 2 & 5-\lambda \\ -2 & -3 \end{vmatrix}\\ &= -\lambda^3 + 5\lambda^2 - 8\lambda + 4,\end{aligned}$$

so $\mathbf{A} - \lambda\mathbf{I}$ is singular if and only if $\lambda^3 - 5\lambda^2 + 8\lambda - 4 = 0$. According to the hint, the integer roots of $p(\lambda) = \lambda^3 - 5\lambda^2 + 8\lambda - 4$ are a subset of $\{\pm 4, \pm 2, \pm 1\}$. Evaluating $p(\lambda)$ at these points reveals that $\lambda = 2$ is a root, and either ordinary or synthetic division produces

$$\frac{p(\lambda)}{\lambda - 2} = \lambda^2 - 3\lambda + 2 = (\lambda - 2)(\lambda - 1).$$

Therefore, $p(\lambda) = (\lambda - 2)^2(\lambda - 1)$, so $\lambda = 2$ and $\lambda = 1$ are the roots of $p(\lambda)$, and these are the values for which $\mathbf{A} - \lambda\mathbf{I}$ is singular.

6.2.23. The indicated substitutions produce the system

$$\begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_{n-1}' \\ x_n' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -p_n & -p_{n-1} & -p_{n-2} & \cdots & -p_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix}.$$

Each of the n vectors $\mathbf{w}_i = \begin{pmatrix} f_i(t) & f_i'(t) & \cdots & f_i^{(n-1)} \end{pmatrix}^T$ for $i = 1, 2, \ldots, n$ satisfies this system, so (6.2.8) may be applied to produce the desired conclusion.

6.2.24. The result is clearly true for $n = 2$. Assume the formula holds for $n = k - 1$, and prove that it must also hold for $n = k$. According to the cofactor expansion in terms of the first row, $\deg p(\lambda) = k - 1$, and it's clear that

$$p(x_2) = p(x_3) = \cdots = p(x_k) = 0,$$

so $x_2, x_3, \ldots, x_k$ are the $k - 1$ roots of $p(\lambda)$. Consequently,

$$p(\lambda) = \alpha(\lambda - x_2)(\lambda - x_3)\cdots(\lambda - x_k),$$

where α is the coefficient of λ^{k-1}. But the coefficient of λ^{k-1} is the cofactor associated with the $(1, k)$-entry, so the induction hypothesis yields

$$\alpha = (-1)^{k-1} \begin{vmatrix} 1 & x_2 & x_2^2 & \cdots & x_2^{k-2} \\ 1 & x_3 & x_3^2 & \cdots & x_3^{k-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_k & x_k^2 & \cdots & x_k^{k-2} \end{vmatrix}_{k-1 \times k-1} = (-1)^{k-1} \prod_{j > i \geq 2} (x_j - x_i).$$

Therefore,

$$\begin{aligned} \det(\mathbf{V}_k) = p(x_1) &= (x_1 - x_2)(x_1 - x_3)\cdots(x_1 - x_k)\alpha \\ &= (x_1 - x_2)(x_1 - x_3)\cdots(x_1 - x_k)\Big((-1)^{k-1} \prod_{j > i \geq 2} (x_j - x_i)\Big) \\ &= (x_2 - x_1)(x_3 - x_1)\cdots(x_k - x_1) \prod_{j > i \geq 2} (x_j - x_i) \\ &= \prod_{j > i} (x_j - x_i), \end{aligned}$$

and the formula is proven. The determinant is nonzero if and only if the x_i's are distinct numbers, and this agrees with the conclusion in Example 4.3.4.

6.2.25. According to (6.1.19),

$$\frac{d\Big(\det(\mathbf{A})\Big)}{dx} = \det(\mathbf{D}_1) + \det(\mathbf{D}_2) + \cdots + \det(\mathbf{D}_n),$$

where $\mathbf{D}_i$ is the matrix

$$\mathbf{D}_i = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{i1}' & a_{i2}' & \cdots & a_{in}' \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Expanding $\det(\mathbf{D}_i)$ in terms of cofactors of the ith row yields

$$\det(\mathbf{A}_i) = a'_{i1}\mathring{A}_{i1} + a'_{i2}\mathring{A}_{i2} + \cdots + a'_{in}\mathring{A}_{in},$$

so the desired conclusion is obtained.

6.2.26. According to (6.1.19),

$$\frac{\partial \det(\mathbf{A})}{\partial a_{ij}} = \det(\mathbf{D}_i) = \begin{vmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{vmatrix} \leftarrow \text{row } i = \mathring{A}_{ij}.$$

6.2.27. The $\binom{4}{2} = 6$ ways to choose pairs of column indices are

$$\begin{array}{ccc} (1,2) & (1,3) & (1,4) \\ & (2,3) & (2,4) \\ & & (3,4) \end{array}$$

so that the Laplace expansion using $i_1 = 1$ and $i_2 = 3$ is

$$\begin{aligned} \det(\mathbf{A}) &= \det \mathbf{A}(1,3\,|\,1,2)\,\mathring{A}(1,3\,|\,1,2) + \det \mathbf{A}(1,3\,|\,1,3)\,\mathring{A}(1,3\,|\,1,3) \\ &\quad + \det \mathbf{A}(1,3\,|\,1,4)\,\mathring{A}(1,3\,|\,1,4) + \det \mathbf{A}(1,3\,|\,2,3)\,\mathring{A}(1,3\,|\,2,3) \\ &\quad + \det \mathbf{A}(1,3\,|\,2,4)\,\mathring{A}(1,3\,|\,2,4) + \det \mathbf{A}(1,3\,|\,3,4)\,\mathring{A}(1,3\,|\,3,4) \\ &= 0 + (-2)(-4) + (-1)(3)(-2) + 0 + (-3)(-3) + (-1)(-8)(2) \\ &= 39. \end{aligned}$$

Solutions for Chapter 7

Solutions for exercises in section 7. 1

7.1.1. $\sigma(\mathbf{A}) = \{-3, 4\}$

$$N(\mathbf{A}+3\mathbf{I}) = span\left\{\begin{pmatrix}-1\\1\end{pmatrix}\right\} \quad \text{and} \quad N(\mathbf{A}-4\mathbf{I}) = span\left\{\begin{pmatrix}-1/2\\1\end{pmatrix}\right\}$$

$\sigma(\mathbf{B}) = \{-2, 2\}$ in which the algebraic multiplicity of $\lambda = -2$ is two.

$$N(\mathbf{B}+2\mathbf{I}) = span\left\{\begin{pmatrix}-4\\1\\0\end{pmatrix}, \begin{pmatrix}-2\\0\\1\end{pmatrix}\right\} \text{ and } N(\mathbf{B}-2\mathbf{I}) = span\left\{\begin{pmatrix}-1/2\\-1/2\\1\end{pmatrix}\right\}$$

$\sigma(\mathbf{C}) = \{3\}$ in which the algebraic multiplicity of $\lambda = 3$ is three.

$$N(\mathbf{C}-3\mathbf{I}) = span\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}\right\}$$

$\sigma(\mathbf{D}) = \{3\}$ in which the algebraic multiplicity of $\lambda = 3$ is three.

$$N(\mathbf{D}-3\mathbf{I}) = span\left\{\begin{pmatrix}2\\1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\1\end{pmatrix}\right\}$$

$\sigma(\mathbf{E}) = \{3\}$ in which the algebraic multiplicity of $\lambda = 3$ is three.

$$N(\mathbf{E}-3\mathbf{I}) = span\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$$

Matrices $\mathbf{C}$ and $\mathbf{D}$ are deficient in eigenvectors.

7.1.2. Form the product $\mathbf{Ax}$, and answer the question, "Is $\mathbf{Ax}$ some multiple of $\mathbf{x}$?" When the answer is *yes,* then $\mathbf{x}$ is an eigenvector for $\mathbf{A}$, and the multiplier is the associated eigenvalue. For this matrix, (a), (c), and (d) are eigenvectors associated with eigenvalues 1, 3, and 3, respectively.

7.1.3. The characteristic polynomial for $\mathbf{T}$ is

$$\det(\mathbf{T}-\lambda\mathbf{I}) = (t_{11}-\lambda)(t_{22}-\lambda)\cdots(t_{nn}-\lambda),$$

so the roots are the t_{ii}'s.

7.1.4. This follows directly from (6.1.16) because

$$\det(\mathbf{T}-\lambda\mathbf{I}) = \begin{vmatrix} \mathbf{A}-\lambda\mathbf{I} & \mathbf{B} \\ \mathbf{0} & \mathbf{C}-\lambda\mathbf{I} \end{vmatrix} = \det(\mathbf{A}-\lambda\mathbf{I})\det(\mathbf{C}-\lambda\mathbf{I}).$$

7.1.5. If λ_i is not repeated, then $N(\mathbf{A}-\lambda_i\mathbf{I}) = span\{\mathbf{e}_i\}$. If the algebraic multiplicity of λ_i is k, and if λ_i occupies positions $i_1, i_2, \ldots, i_k$ in $\mathbf{D}$, then

$$N(\mathbf{A}-\lambda_i\mathbf{I}) = span\{\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \ldots, \mathbf{e}_{i_k}\}.$$

7.1.6. $\mathbf{A}$ singular $\Longleftrightarrow \det(\mathbf{A}) = 0 \Longleftrightarrow 0$ solves $\det(\mathbf{A}-\lambda\mathbf{I}) = 0 \Longleftrightarrow 0 \in \sigma(\mathbf{A})$.

7.1.7. Zero is not in or on any Gerschgorin circle. You could also say that $\mathbf{A}$ is nonsingular because it is diagonally dominant—see Example 7.1.6 on p. 499.

7.1.8. If $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}^*\mathbf{A}$, then $\|\mathbf{A}\mathbf{x}\|_2^2 / \|\mathbf{x}\|_2^2 = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x}/\mathbf{x}^*\mathbf{x} = \lambda$ is real and nonnegative. Furthermore, $\lambda > 0$ if and only if $\mathbf{A}^*\mathbf{A}$ is nonsingular or, equivalently, $n = rank(\mathbf{A}^*\mathbf{A}) = rank(\mathbf{A})$. Similar arguments apply to $\mathbf{A}\mathbf{A}^*$.

7.1.9. (a) $\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \implies \mathbf{x} = \lambda\mathbf{A}^{-1}\mathbf{x} \implies (1/\lambda)\mathbf{x} = \mathbf{A}^{-1}\mathbf{x}$.

(b) $\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \Longleftrightarrow (\mathbf{A}-\alpha\mathbf{I})\mathbf{x} = (\lambda-\alpha)\mathbf{x} \Longleftrightarrow (\lambda-\alpha)^{-1}\mathbf{x} = (\mathbf{A}-\alpha\mathbf{I})^{-1}\mathbf{x}$.

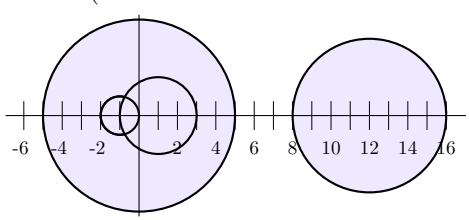
7.1.10. (a) Successively use $\mathbf{A}$ as a left-hand multiplier to produce

$$\begin{aligned} \mathbf{A}\mathbf{x} = \lambda\mathbf{x} &\implies \mathbf{A}^2\mathbf{x} = \lambda\mathbf{A}\mathbf{x} = \lambda^2\mathbf{x} \\ &\implies \mathbf{A}^3\mathbf{x} = \lambda^2\mathbf{A}\mathbf{x} = \lambda^3\mathbf{x} \\ &\implies \mathbf{A}^4\mathbf{x} = \lambda^3\mathbf{A}\mathbf{x} = \lambda^4\mathbf{x} \\ &\quad \text{etc.} \end{aligned}$$

(b) Use part (a) to write

$$p(\mathbf{A})\mathbf{x} = \left(\sum_i \alpha_i\mathbf{A}^i\right)\mathbf{x} = \sum_i \alpha_i\mathbf{A}^i\mathbf{x} = \sum_i \alpha_i\lambda^i\mathbf{x} = \left(\sum_i \alpha_i\lambda^i\right)\mathbf{x} = p(\lambda)\mathbf{x}.$$

7.1.11. Since one Geschgorin circle (derived from row sums and shown below) is isolated



from the union of the other three circles, statement (7.1.14) on p. 498 insures that there is one eigenvalue in the isolated circle and three eigenvalues in the union of the other three. But, as discussed on p. 492, the eigenvalues of real matrices occur in conjugate pairs. So, the root in the isolated circle must be real and there must be at least one real root in the union of the other three circles. Computation reveals that $\sigma(\mathbf{A}) = \{\pm\mathrm{i}, 2, 10\}$.

7.1.12. Use Exercise 7.1.10 to deduce that

$$\lambda \in \sigma(\mathbf{A}) \implies \lambda^k \in \sigma\left(\mathbf{A}^k\right) \implies \lambda^k = 0 \implies \lambda = 0.$$

Therefore, (7.1.7) insures that $trace(\mathbf{A}) = \sum_i \lambda_i = 0$.

7.1.13. This is true because $N(\mathbf{A} - \lambda\mathbf{I})$ is a subspace—recall that subspaces are closed under vector addition and scalar multiplication.

7.1.14. If there exists a nonzero vector $\mathbf{x}$ that satisfies $\mathbf{A}\mathbf{x} = \lambda_1\mathbf{x}$ and $\mathbf{A}\mathbf{x} = \lambda_2\mathbf{x}$, where $\lambda_1 \neq \lambda_2$, then

$$\mathbf{0} = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x} = \lambda_1\mathbf{x} - \lambda_2\mathbf{x} = (\lambda_1 - \lambda_2)\mathbf{x}.$$

But this implies $\mathbf{x} = \mathbf{0}$, which is impossible. Consequently, no such $\mathbf{x}$ can exist.

7.1.15. No—consider $\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$.

7.1.16. Almost any example with rather random entries will do the job, but avoid diagonal or triangular matrices—they are too special.

7.1.17. (a) $\mathbf{c} = (\mathbf{A} - \lambda\mathbf{I})^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{c} = (\mathbf{A} - \lambda\mathbf{I})^{-1}(\mathbf{A}\mathbf{c} - \lambda\mathbf{c}) = (\mathbf{A} - \lambda\mathbf{I})^{-1}(\lambda_k - \lambda)\mathbf{c}$.

(b) Use (6.2.3) to compute the characteristic polynomial for $\mathbf{A} + \mathbf{c}\mathbf{d}^T$ to be

$$\begin{aligned}
\det\left(\mathbf{A} + \mathbf{c}\mathbf{d}^T - \lambda\mathbf{I}\right) &= \det\left(\mathbf{A} - \lambda\mathbf{I} + \mathbf{c}\mathbf{d}^T\right) \\
&= \det(\mathbf{A} - \lambda\mathbf{I})\left(1 + \mathbf{d}^T(\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{c}\right) \\
&= \left(\pm\prod_{i=1}^{n}(\lambda_j - \lambda)\right)\left(1 + \frac{\mathbf{d}^T\mathbf{c}}{\lambda_k - \lambda}\right) \\
&= \left(\pm\prod_{j\neq k}(\lambda_j - \lambda)\right)\left(\lambda_k + \mathbf{d}^T\mathbf{c} - \lambda\right).
\end{aligned}$$

The roots of this polynomial are $\lambda_1, \ldots, \lambda_{k-1}, \lambda_k + \mathbf{d}^T\mathbf{c}, \lambda_{k+1}, \ldots, \lambda_n$.

(c) $\mathbf{d} = \dfrac{(\mu - \lambda_k)\mathbf{c}}{\mathbf{c}^T\mathbf{c}}$ will do the job.

7.1.18. (a) The transpose does not alter the determinant—recall (6.1.4)—so that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det\left(\mathbf{A}^T - \lambda\mathbf{I}\right).$$

(b) We know from Exercise 6.1.8 that $\overline{\det(\mathbf{A})} = \det(\mathbf{A}^*)$, so

$$\begin{aligned}\lambda \in \sigma(\mathbf{A}) &\iff 0 = \det(\mathbf{A} - \lambda\mathbf{I}) \\ &\iff 0 = \overline{\det(\mathbf{A} - \lambda\mathbf{I})} = \det((\mathbf{A} - \lambda\mathbf{I})^*) = \det\left(\mathbf{A}^* - \overline{\lambda}\mathbf{I}\right) \\ &\iff \overline{\lambda} \in \sigma(\mathbf{A}^*).\end{aligned}$$

(c) Yes.

(d) Apply the reverse order law for conjugate transposes to obtain

$$\mathbf{y}^*\mathbf{A} = \mu\mathbf{y}^* \implies \mathbf{A}^*\mathbf{y} = \overline{\mu}\mathbf{y} \implies \mathbf{A}^T\mathbf{y} = \overline{\mu}\mathbf{y} \implies \overline{\mu} \in \sigma\left(\mathbf{A}^T\right) = \sigma(\mathbf{A}),$$

and use the conclusion of part (c) insuring that the eigenvalues of real matrices must occur in conjugate pairs.

7.1.19. (a) When $m = n$, Exercise 6.2.16 insures that

$$\lambda^n \det(\mathbf{AB} - \lambda\mathbf{I}) = \lambda^n \det(\mathbf{BA} - \lambda\mathbf{I}) \quad \text{for all } \lambda,$$

so $\det(\mathbf{AB} - \lambda\mathbf{I}) = \det(\mathbf{BA} - \lambda\mathbf{I})$.

(b) If $m \neq n$, then the characteristic polynomials of $\mathbf{AB}$ and $\mathbf{BA}$ are of degrees m and n, respectively, so they must be different. When m and n are different—say $m > n$—Exercise 6.2.16 implies that

$$\det(\mathbf{AB} - \lambda\mathbf{I}) = (-\lambda)^{m-n} \det(\mathbf{BA} - \lambda\mathbf{I}).$$

Consequently, $\mathbf{AB}$ has $m - n$ more zero eigenvalues than $\mathbf{BA}$.

7.1.20. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$, and suppose $\mathbf{X}$ is $n \times g$. The equation $(\mathbf{A} - \lambda\mathbf{I})\mathbf{BX} = \mathbf{0}$ says that the columns of $\mathbf{BX}$ are in $N(\mathbf{A} - \lambda\mathbf{I})$, and hence they are linear combinations of the basis vectors in $\mathbf{X}$. Thus

$$[\mathbf{BX}]_{*j} = \sum_i p_{ij}\mathbf{X}_{*j} \implies \mathbf{BX} = \mathbf{XP}, \quad \text{where } \mathbf{P}_{g\times g} = [p_{ij}].$$

If $(\mu, \mathbf{z})$ is any eigenpair for $\mathbf{P}$, then

$$\mathbf{B}(\mathbf{Xz}) = \mathbf{XPz} = \mu(\mathbf{Xz}) \quad \text{and} \quad \mathbf{AX} = \lambda\mathbf{X} \implies \mathbf{A}(\mathbf{Xz}) = \lambda(\mathbf{Xz}),$$

so $\mathbf{Xz}$ is a common eigenvector.

7.1.21. (a) If $\mathbf{Px} = \lambda\mathbf{x}$ and $\mathbf{y}^*\mathbf{Q} = \mu\mathbf{y}^*$, then $\mathbf{T}(xy^*) = \mathbf{Pxy}^*\mathbf{Q} = \lambda\mu\mathbf{xy}^*$.

(b) Since $\dim \mathcal{C}^{m\times n} = mn$, the operator $\mathbf{T}$ (as well as any coordinate matrix representation of $\mathbf{T}$) must have exactly mn eigenvalues (counting multiplicities), and since there are exactly mn products $\lambda\mu$, where $\lambda \in \sigma(\mathbf{P})$, $\mu \in \sigma(\mathbf{Q})$, it follows that $\sigma(\mathbf{T}) = \{\lambda\mu \mid \lambda \in \sigma(\mathbf{P}), \mu \in \sigma(\mathbf{Q})\}$. Use the fact

that the trace is the sum of the eigenvalues (recall (7.1.7)) to conclude that $trace\,(\mathbf{T}) = \sum_{i,j} \lambda_i \mu_j = \sum_i \lambda_i \sum_j \mu_j = trace\,(\mathbf{P})\, trace\,(\mathbf{Q})$.

7.1.22. (a) Use (6.2.3) to compute the characteristic polynomial for $\mathbf{D} + \alpha \mathbf{v}\mathbf{v}^T$ to be

$$\begin{aligned} p(\lambda) &= \det\left(\mathbf{D} + \alpha\mathbf{v}\mathbf{v}^T - \lambda\mathbf{I}\right) \\ &= \det\left(\mathbf{D} - \lambda\mathbf{I} + \alpha\mathbf{v}\mathbf{v}^T\right) \\ &= \det\left(\mathbf{D} - \lambda\mathbf{I}\right)\left(1 + \alpha\mathbf{v}^T(\mathbf{D} - \lambda\mathbf{I})^{-1}\mathbf{v}\right) \qquad (\ddagger) \\ &= \left(\prod_{j=1}^{n}(\lambda - \lambda_j)\right)\left(1 + \alpha\sum_{i=1}^{n}\frac{v_i^2}{\lambda_i - \lambda}\right) \\ &= \prod_{j=1}^{n}(\lambda - \lambda_j) + \alpha\sum_{i=1}^{n}\left(v_i\prod_{j\neq i}(\lambda - \lambda_j)\right). \end{aligned}$$

For each λ_k, it is true that

$$p(\lambda_k) = \alpha v_k \prod_{j\neq k}(\lambda_k - \lambda_j) \neq 0,$$

and hence no λ_k can be an eigenvalue for $\mathbf{D} + \alpha\mathbf{v}\mathbf{v}^T$. Consequently, if ξ is an eigenvalue for $\mathbf{D} + \alpha\mathbf{v}\mathbf{v}^T$, then $\det(\mathbf{D} - \xi\mathbf{I}) \neq 0$, so $p(\xi) = 0$ and $(\ddagger)$ imply that

$$0 = 1 + \alpha\mathbf{v}^T(\mathbf{D} - \xi\mathbf{I})^{-1}\mathbf{v} = 1 + \alpha\sum_{i=1}^{n}\frac{v_i^2}{\lambda_i - \xi} = f(\xi).$$

(b) Use the fact that $f(\xi_i) = 1 + \alpha\mathbf{v}^T(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v} = 0$ to write

$$\begin{aligned} \left(\mathbf{D} + \alpha\mathbf{v}\mathbf{v}^T\right)(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v} &= \mathbf{D}(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v} + \mathbf{v}\left(\alpha\mathbf{v}^T(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v}\right) \\ &= \mathbf{D}(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v} - \mathbf{v} \\ &= \Big(\mathbf{D} - (\mathbf{D} - \xi_i\mathbf{I})\Big)(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v} \\ &= \xi_i(\mathbf{D} - \xi_i\mathbf{I})^{-1}\mathbf{v}. \end{aligned}$$

7.1.23. (a) If $p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$, then

$$\ln p(\lambda) = \sum_{i=1}^{n}\ln(\lambda - \lambda_i) \quad\Longrightarrow\quad \frac{p'(\lambda)}{p(\lambda)} = \sum_{i=1}^{n}\frac{1}{(\lambda - \lambda_i)}.$$

(b) If $|\lambda_i/\lambda| < 1$, then we can write

$$(\lambda - \lambda_i)^{-1} = \left(\lambda\left(1 - \frac{\lambda_i}{\lambda}\right)\right)^{-1} = \frac{1}{\lambda}\left(1 - \frac{\lambda_i}{\lambda}\right)^{-1} = \frac{1}{\lambda}\left(1 + \frac{\lambda_i}{\lambda} + \frac{\lambda_i^2}{\lambda^2} + \cdots\right).$$

Consequently,

$$\sum_{i=1}^{n} \frac{1}{(\lambda - \lambda_i)} = \sum_{i=1}^{n} \left(\frac{1}{\lambda} + \frac{\lambda_i}{\lambda^2} + \frac{\lambda_i^2}{\lambda^3} + \cdots \right) = \frac{n}{\lambda} + \frac{\tau_1}{\lambda^2} + \frac{\tau_2}{\lambda^3} + \cdots .$$

(c) Combining these two results yields

$$\begin{aligned}
&n\lambda^{n-1} + (n-1)c_1\lambda^{n-2} + (n-2)c_2\lambda^{n-3} + \cdots + c_{n-1} \\
&= \left(\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_n\right)\left(\frac{n}{\lambda} + \frac{\tau_1}{\lambda^2} + \frac{\tau_2}{\lambda^3} + \cdots\right) \\
&= n\lambda^{n-1} + (nc_1 + \tau_1)\,\lambda^{n-2} + (nc_2 + \tau_1 c_1 + \tau_2)\,\lambda^{n-3} \\
&\qquad + \cdots + (nc_{n-1} + \tau_1 c_{n-2} + \tau_2 c_{n-3} + \cdots + \tau_{n-1}) \\
&\qquad + (nc_n + \tau_1 c_{n-1} + \tau_2 c_{n-2} \cdots + \tau_n)\,\frac{1}{\lambda} + \cdots ,
\end{aligned}$$

and equating like powers of λ produces the desired conclusion.

7.1.24. We know from Exercise 7.1.10 that $\lambda \in \sigma(\mathbf{A}) \implies \lambda^k \in \sigma\left(\mathbf{A}^k\right)$, so (7.1.7) guarantees that $trace\left(\mathbf{A}^k\right) = \sum_i \lambda_i^k = \tau_k$. Proceed by induction. The result is true for $k = 1$ because (7.1.7) says that $c_1 = -trace\,(\mathbf{A})$. Assume that

$$c_i = -\frac{trace\,(\mathbf{A}\mathbf{B}_{i-1})}{i} \quad \text{for} \quad i = 1, 2, \ldots, k-1,$$

and prove the result holds for $i = k$. Recursive application of the induction hypothesis produces

$$\begin{aligned}
\mathbf{B}_1 &= c_1\mathbf{I} + \mathbf{A} \\
\mathbf{B}_2 &= c_2\mathbf{I} + c_1\mathbf{A} + \mathbf{A}^2 \\
&\vdots \\
\mathbf{B}_{k-1} &= c_{k-1}\mathbf{I} + c_{k-2}\mathbf{A} + \cdots + c_1\mathbf{A}^{k-2} + \mathbf{A}^{k-1},
\end{aligned}$$

and therefore we can use Newton's identities given in Exercise 7.1.23 to obtain

$$\begin{aligned}
trace\,(\mathbf{A}\mathbf{B}_{k-1}) &= trace\left(c_{k-1}\mathbf{A} + c_{k-2}\mathbf{A}^2 + \cdots + c_1\mathbf{A}^{k-1} + \mathbf{A}^k\right) \\
&= c_{k-1}\tau_1 + c_{k-2}\tau_2 + \cdots + c_1\tau_{k-1} + \tau_k \\
&= -kc_k.
\end{aligned}$$

Solutions for exercises in section 7. 2

7.2.1. The characteristic equation is $\lambda^2-2\lambda-8=(\lambda+2)(\lambda-4)=0$, so the eigenvalues are $\lambda_1=-2$ and $\lambda_2=4$. Since no eigenvalue is repeated, (7.2.6) insures $\mathbf{A}$ must be diagonalizable. A similarity transformation $\mathbf{P}$ that diagonalizes $\mathbf{A}$ is constructed from a complete set of independent eigenvectors. Compute a pair of eigenvectors associated with λ_1 and λ_2 to be

$$\mathbf{x}_1=\begin{pmatrix}-1\\1\end{pmatrix},\ \mathbf{x}_2=\begin{pmatrix}-1\\2\end{pmatrix},\quad \text{and set}\quad \mathbf{P}=\begin{pmatrix}-1&-1\\1&2\end{pmatrix}.$$

Now verify that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\begin{pmatrix}-2&-1\\1&1\end{pmatrix}\begin{pmatrix}-8&-6\\12&10\end{pmatrix}\begin{pmatrix}-1&-1\\1&2\end{pmatrix}=\begin{pmatrix}-2&0\\0&4\end{pmatrix}=\mathbf{D}.$$

7.2.2. (a) The characteristic equation is $\lambda^3-3\lambda-2=(\lambda-2)(\lambda+1)^2=0$, so the eigenvalues are $\lambda=2$ and $\lambda=-1$. By reducing $\mathbf{A}-2\mathbf{I}$ and $\mathbf{A}+\mathbf{I}$ to echelon form, compute bases for $N(\mathbf{A}-2\mathbf{I})$ and $N(\mathbf{A}+\mathbf{I})$. One set of bases is

$$N(\mathbf{A}-2\mathbf{I})=span\left\{\begin{pmatrix}-1\\0\\2\end{pmatrix}\right\}\quad\text{and}\quad N(\mathbf{A}+\mathbf{I})=span\left\{\begin{pmatrix}-1\\1\\0\end{pmatrix},\begin{pmatrix}-1\\0\\1\end{pmatrix}\right\}.$$

Therefore,

$$\begin{aligned}geo\ mult_{\mathbf{A}}(2)&=\dim N(\mathbf{A}-2\mathbf{I})=1=alg\ mult_{\mathbf{A}}(2),\\ geo\ mult_{\mathbf{A}}(-1)&=\dim N(\mathbf{A}+\mathbf{I})=2=alg\ mult_{\mathbf{A}}(-1).\end{aligned}$$

In other words, $\lambda=2$ is a simple eigenvalue, and $\lambda=-1$ is a semisimple eigenvalue.

(b) A similarity transformation $\mathbf{P}$ that diagonalizes $\mathbf{A}$ is constructed from a complete set of independent eigenvectors, and these are obtained from the above bases. Set $\mathbf{P}=\begin{pmatrix}-1&-1&-1\\0&1&0\\2&0&1\end{pmatrix}$, and compute $\mathbf{P}^{-1}=\begin{pmatrix}1&1&1\\0&1&0\\-2&-2&-1\end{pmatrix}$ and verify that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}=\begin{pmatrix}2&0&0\\0&-1&0\\0&0&-1\end{pmatrix}$.

7.2.3. Consider the matrix $\mathbf{A}$ of Exercise 7.2.1. We know from its solution that $\mathbf{A}$ is similar to $\mathbf{D}=\begin{pmatrix}-2&0\\0&4\end{pmatrix}$, but the two eigenspaces for $\mathbf{A}$ are spanned by $\begin{pmatrix}-1\\1\end{pmatrix}$ and $\begin{pmatrix}-1\\2\end{pmatrix}$, whereas the eigenspaces for $\mathbf{D}$ are spanned by the unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$.

7.2.4. The characteristic equation of $\mathbf{A}$ is $p(\lambda) = (\lambda-1)(\lambda-2)^2$, so $alg\ mult_{\mathbf{A}}(2) = 2$. To find $geo\ mult_{\mathbf{A}}(2)$, reduce $\mathbf{A} - 2\mathbf{I}$ to echelon form to find that

$$N(\mathbf{A} - 2\mathbf{I}) = span\left\{\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}\right\},$$

so $geo\ mult_{\mathbf{A}}(2) = \dim N(\mathbf{A} - 2\mathbf{I}) = 1$. Since there exists at least one eigenvalue such that $geo\ mult_{\mathbf{A}}(\lambda) < alg\ mult_{\mathbf{A}}(\lambda)$, it follows (7.2.5) on p. 512 that $\mathbf{A}$ cannot be diagonalized by a similarity transformation.

7.2.5. A formal induction argument can be given, but it suffices to "do it with dots" by writing

$$\begin{aligned}\mathbf{B}^k &= (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})(\mathbf{P}^{-1}\mathbf{A}\mathbf{P})\cdots(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}) \\ &= \mathbf{P}^{-1}\mathbf{A}(\mathbf{P}\mathbf{P}^{-1})\mathbf{A}(\mathbf{P}\mathbf{P}^{-1})\cdots(\mathbf{P}\mathbf{P}^{-1})\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}\mathbf{A}\cdots\mathbf{A}\mathbf{P} = \mathbf{P}^{-1}\mathbf{A}^k\mathbf{P}.\end{aligned}$$

7.2.6. $\lim_{n\to\infty}\mathbf{A}^n = \begin{pmatrix} 5 & 2 \\ -10 & -4 \end{pmatrix}$. Of course, you could compute $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^3, \ldots$ in hopes of seeing a pattern, but this clumsy approach is not definitive. A better technique is to diagonalize $\mathbf{A}$ with a similarity transformation, and then use the result of Exercise 7.2.5. The characteristic equation is $0 = \lambda^2-(19/10)\lambda+(1/2) = (\lambda-1)(\lambda-(9/10))$, so the eigenvalues are $\lambda = 1$ and $\lambda = .9$. By reducing $\mathbf{A}-\mathbf{I}$ and $\mathbf{A} - .9\mathbf{I}$ to echelon form, we see that

$$N(\mathbf{A} - \mathbf{I}) = span\left\{\begin{pmatrix} -1 \\ 2 \end{pmatrix}\right\} \quad \text{and} \quad N(\mathbf{A} - .9\mathbf{I}) = span\left\{\begin{pmatrix} -2 \\ 5 \end{pmatrix}\right\},$$

so $\mathbf{A}$ is indeed diagonalizable, and $\mathbf{P} = \begin{pmatrix} -1 & -2 \\ 2 & 5 \end{pmatrix}$ is a matrix such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & .9 \end{pmatrix} = \mathbf{D}$ or, equivalently, $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$. The result of Exercise 7.2.5 says that $\mathbf{A}^n = \mathbf{P}\mathbf{D}^n\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & .9^n \end{pmatrix}\mathbf{P}^{-1}$, so

$$\lim_{n\to\infty}\mathbf{A}^n = \mathbf{P}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{P}^{-1} = \begin{pmatrix} -1 & -2 \\ 2 & 5 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} -5 & -2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ -10 & -4 \end{pmatrix}.$$

7.2.7. It follows from $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$ that $\mathbf{y}_i^*\mathbf{x}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$ as well as $\mathbf{y}_i^*\mathbf{X} = \mathbf{0}$ and $\mathbf{Y}^*\mathbf{x}_i = \mathbf{0}$ for each $i = 1, \ldots, t$, so

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_t^* \\ \mathbf{Y}^* \end{pmatrix}\mathbf{A}\Big(\mathbf{x}_1 \,|\, \cdots \,|\, \mathbf{x}_t \,|\, \mathbf{X}\Big) = \begin{pmatrix} \lambda_1 & \cdots & 0 & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \lambda_t & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{Y}^*\mathbf{A}\mathbf{X} \end{pmatrix} = \mathbf{B}.$$

Therefore, examining the first t rows on both sides of $\mathbf{P}^{-1}\mathbf{A} = \mathbf{B}\mathbf{P}^{-1}$ yields $\mathbf{y}_i^*\mathbf{A} = \lambda_i \mathbf{y}_i^*$ for $i = 1, \ldots, t$.

7.2.8. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then $\mathbf{P}^{-1}\mathbf{A}^k\mathbf{P} = \operatorname{diag}\left(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\right)$ for $k = 0, 1, 2, \ldots$ or, equivalently, $\mathbf{A}^k = \mathbf{P}\operatorname{diag}\left(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\right)\mathbf{P}^{-1}$. Therefore, $\mathbf{A}^k \to \mathbf{0}$ if and only if each $\lambda_i^k \to 0$, which is equivalent to saying that $|\lambda_i| < 1$ for each i. Since $\rho(\mathbf{A}) = \max_{\lambda_i \in \sigma(\mathbf{A})} |\lambda_i|$ (recall Example 7.1.4 on p. 497), it follows that $\mathbf{A}^k \to \mathbf{0}$ if and only if $\rho(\mathbf{A}) < 1$.

7.2.9. The characteristic equation for $\mathbf{A}$ is $\lambda^2 - 2\lambda + 1$, so $\lambda = 1$ is the only distinct eigenvalue. By reducing $\mathbf{A} - \mathbf{I}$ to echelon form, we see that $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ is a basis for $N(\mathbf{A} - \mathbf{I})$, so $\mathbf{x} = (1/5)\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ is an eigenvector of unit length. Following the procedure on p. 325, we find that $\mathbf{R} = \begin{pmatrix} 3/5 & 4/5 \\ 4/5 & -3/5 \end{pmatrix}$ is an elementary reflector having $\mathbf{x}$ as its first column, and $\mathbf{R}^T\mathbf{A}\mathbf{R} = \mathbf{R}\mathbf{A}\mathbf{R} = \begin{pmatrix} 1 & 25 \\ 0 & 1 \end{pmatrix}$.

7.2.10. From Example 7.2.1 on p. 507 we see that the characteristic equation for $\mathbf{A}$ is $p(\lambda) = \lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2 = 0$. Straightforward computation shows that

$$p(\mathbf{A}) = (\mathbf{A} - \mathbf{I})(\mathbf{A} + 3\mathbf{I})^2 = \begin{pmatrix} 0 & -4 & -4 \\ 8 & -12 & -8 \\ -8 & 8 & 4 \end{pmatrix}\begin{pmatrix} 16 & -16 & -16 \\ 32 & -32 & -32 \\ -32 & 32 & 32 \end{pmatrix} = \mathbf{0}.$$

7.2.11. Rescale the observed eigenvector as $\mathbf{x} = (1/2)(1, 1, 1, 1)^T = \mathbf{y}$ so that $\mathbf{x}^T\mathbf{x} = 1$. Follow the procedure described in Example 5.6.3 (p. 325), and set $\mathbf{u} = \mathbf{x} - \mathbf{e}_1$ to construct

$$\mathbf{R} = \mathbf{I} - \frac{2\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} = \mathbf{P} = \left[\mathbf{x} \,|\, \mathbf{X}\right] \quad (\text{since } \mathbf{x} = \mathbf{y}).$$

Consequently, $\mathbf{B} = \mathbf{X}^T\mathbf{A}\mathbf{X} = \begin{pmatrix} -1 & 0 & -1 \\ 0 & 2 & 0 \\ -1 & 0 & 1 \end{pmatrix}$, and $\sigma(\mathbf{B}) = \{2, \sqrt{2}, -\sqrt{2}\}$.

7.2.12. Use the spectral theorem with properties $\mathbf{G}_i\mathbf{G}_j = \mathbf{0}$ for $i \neq j$ and $\mathbf{G}_i^2 = \mathbf{G}_i$ to write $\mathbf{A}\mathbf{G}_i = (\lambda_1\mathbf{G}_1 + \lambda_2\mathbf{G}_2 + \cdots + \lambda_k\mathbf{G}_k)\mathbf{G}_i = \lambda_i\mathbf{G}_i^2 = \lambda_i\mathbf{G}_i$. A similar argument shows $\mathbf{G}_i\mathbf{A} = \lambda_i\mathbf{G}_i$.

7.2.13. Use (6.2.3) to show that $\lambda^{n-1}(\lambda - \mathbf{d}^T\mathbf{c}) = 0$ is the characteristic equation for $\mathbf{A}$. Thus $\lambda = 0$ and $\lambda = \mathbf{d}^T\mathbf{c}$ are the eigenvalues of $\mathbf{A}$. We know from (7.2.5) that $\mathbf{A}$ is diagonalizable if and only if the algebraic and geometric multiplicities agree for each eigenvalue. Since $\mathit{geo\ mult}_{\mathbf{A}}(0) = \dim N(\mathbf{A}) = n - \mathit{rank}(\mathbf{A}) = n - 1$, and since

$$\mathit{alg\ mult}_{\mathbf{A}}(0) = \begin{cases} n - 1 & \text{if} \quad \mathbf{d}^T\mathbf{c} \neq 0, \\ n & \text{if} \quad \mathbf{d}^T\mathbf{c} = 0, \end{cases}$$

it follows that $\mathbf{A}$ is diagonalizable if and only if $\mathbf{d}^T\mathbf{c} \neq 0$.

7.2.14. If $\mathbf{W}$ and $\mathbf{Z}$ are diagonalizable—say $\mathbf{P}^{-1}\mathbf{W}\mathbf{P}$ and $\mathbf{Q}^{-1}\mathbf{Z}\mathbf{Q}$ are diagonal—then $\begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix}$ diagonalizes $\mathbf{A}$. Use an indirect argument for the converse. Suppose $\mathbf{A}$ is diagonalizable but $\mathbf{W}$ (or $\mathbf{Z}$) is not. Then there is an eigenvalue $\lambda \in \sigma(\mathbf{W})$ with $geo\ mult_{\mathbf{W}}(\lambda) < alg\ mult_{\mathbf{W}}(\lambda)$. Since $\sigma(\mathbf{A}) = \sigma(\mathbf{W}) \cup \sigma(\mathbf{Z})$ (Exercise 7.1.4), this would mean that

$$\begin{aligned} geo\ mult_{\mathbf{A}}(\lambda) &= \dim N(\mathbf{A} - \lambda\mathbf{I}) = (s+t) - rank(\mathbf{A} - \lambda\mathbf{I}) \\ &= (s - rank(\mathbf{W} - \lambda\mathbf{I})) + (t - rank(\mathbf{Z} - \lambda\mathbf{I})) \\ &= \dim N(\mathbf{W} - \lambda\mathbf{I}) + \dim N(\mathbf{Z} - \lambda\mathbf{I}) \\ &= geo\ mult_{\mathbf{W}}(\lambda) + geo\ mult_{\mathbf{Z}}(\lambda) \\ &< alg\ mult_{\mathbf{W}}(\lambda) + alg\ mult_{\mathbf{Z}}(\lambda) \\ &< alg\ mult_{\mathbf{A}}(\lambda), \end{aligned}$$

which contradicts the fact that $\mathbf{A}$ is diagonalizable.

7.2.15. If $\mathbf{AB} = \mathbf{BA}$, then, by Exercise 7.1.20 (p. 503), $\mathbf{A}$ and $\mathbf{B}$ have a common eigenvector—say $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and $\mathbf{B}\mathbf{x} = \mu\mathbf{x}$, where $\mathbf{x}$ has been scaled so that $\|\mathbf{x}\|_2 = 1$. If $\mathbf{R} = \left[\mathbf{x} \,|\, \mathbf{X}\right]$ is a unitary matrix having $\mathbf{x}$ as its first column (Example 5.6.3, p. 325), then

$$\mathbf{R}^*\mathbf{A}\mathbf{R} = \begin{pmatrix} \lambda & \mathbf{x}^*\mathbf{A}\mathbf{X} \\ \mathbf{0} & \mathbf{X}^*\mathbf{A}\mathbf{X} \end{pmatrix} \quad \text{and} \quad \mathbf{R}^*\mathbf{B}\mathbf{R} = \begin{pmatrix} \mu & \mathbf{x}^*\mathbf{B}\mathbf{X} \\ \mathbf{0} & \mathbf{X}^*\mathbf{B}\mathbf{X} \end{pmatrix}.$$

Since $\mathbf{A}$ and $\mathbf{B}$ commute, so do $\mathbf{R}^*\mathbf{A}\mathbf{R}$ and $\mathbf{R}^*\mathbf{B}\mathbf{R}$, which in turn implies $\mathbf{A}_2 = \mathbf{X}^*\mathbf{A}\mathbf{X}$ and $\mathbf{B}_2 = \mathbf{X}^*\mathbf{B}\mathbf{X}$ commute. Thus the problem is deflated, so the same argument can be applied inductively in a manner similar to the development of Schur's triangularization theorem (p. 508).

7.2.16. If $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}_1$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \mathbf{D}_2$ are both diagonal, then $\mathbf{D}_1\mathbf{D}_2 = \mathbf{D}_2\mathbf{D}_1$ implies that $\mathbf{AB} = \mathbf{BA}$. Conversely, suppose $\mathbf{AB} = \mathbf{BA}$. Let $\lambda \in \sigma(\mathbf{A})$ with $alg\ mult_{\mathbf{A}}(\lambda) = a$, and let $\mathbf{P}$ be such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} \lambda\mathbf{I}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}$, where $\mathbf{D}$ is a diagonal matrix with $\lambda \notin \sigma(\mathbf{D})$. Since $\mathbf{A}$ and $\mathbf{B}$ commute, so do $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ and $\mathbf{P}^{-1}\mathbf{B}\mathbf{P}$. Consequently, if $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}$, then

$$\begin{pmatrix} \lambda\mathbf{I}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix}\begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix} = \begin{pmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix}\begin{pmatrix} \lambda\mathbf{I}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \implies \begin{cases} \lambda\mathbf{X} = \mathbf{X}\mathbf{D}, \\ \mathbf{D}\mathbf{Y} = \lambda\mathbf{Y}, \end{cases}$$

so $(\mathbf{D} - \lambda\mathbf{I})\mathbf{X} = \mathbf{0}$ and $(\mathbf{D} - \lambda\mathbf{I})\mathbf{Y} = \mathbf{0}$. But $(\mathbf{D} - \lambda\mathbf{I})$ is nonsingular, so $\mathbf{X} = \mathbf{0}$ and $\mathbf{Y} = \mathbf{0}$, and thus $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$. Since $\mathbf{B}$ is diagonalizable, so is

$\mathbf{P}^{-1}\mathbf{BP}$, and hence so are $\mathbf{W}$ and $\mathbf{Z}$ (Exercise 7.2.14). If $\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_w & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_z \end{pmatrix}$, where $\mathbf{Q}_w$ and $\mathbf{Q}_z$ are such that $\mathbf{Q}_w^{-1}\mathbf{W}\mathbf{Q}_w = \mathbf{D}_w$ and $\mathbf{Q}_z^{-1}\mathbf{Z}\mathbf{Q}_z = \mathbf{D}_z$ are each diagonal, then

$$(\mathbf{PQ})^{-1}\mathbf{A}(\mathbf{PQ}) = \begin{pmatrix} \lambda\mathbf{I}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_z^{-1}\mathbf{D}\mathbf{Q}_z \end{pmatrix} \quad \text{and} \quad (\mathbf{PQ})^{-1}\mathbf{B}(\mathbf{PQ}) = \begin{pmatrix} \mathbf{D}_w & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_z \end{pmatrix}.$$

Thus the problem is deflated because $\mathbf{A}_2 = \mathbf{Q}_z^{-1}\mathbf{D}\mathbf{Q}_z$ and $\mathbf{B}_2 = \mathbf{D}_z$ commute and are diagonalizable, so the same argument can be applied to them. If $\mathbf{A}$ has k distinct eigenvalues, then the desired conclusion is attained after k repetitions.

7.2.17. It's not legitimate to equate $p(\mathbf{A})$ with $\det(\mathbf{A} - \mathbf{AI})$ because the former is a matrix while the latter is a scalar.

7.2.18. This follows from the eigenvalue formula developed in Example 7.2.5 (p. 514) by using the identity $1 - \cos\theta = 2\sin^2(\theta/2)$.

7.2.19. (a) The result in Example 7.2.5 (p. 514) shows that the eigenvalues of $\mathbf{N}+\mathbf{N}^T$ and $\mathbf{N}-\mathbf{N}^T$ are $\lambda_j = 2\cos(j\pi/n+1)$ and $\lambda_j = 2\mathrm{i}\cos(j\pi/n+1)$, respectively.

(b) Since $\mathbf{N} - \mathbf{N}^T$ is skew symmetric, it follows from Exercise 6.1.12 (p. 473) that $\mathbf{N}-\mathbf{N}^T$ is nonsingular if and only if n is even, which is equivalent to saying $\mathbf{N} - \mathbf{N}^T$ has no zero eigenvalues (recall Exercise 7.1.6, p. 501), and hence, by part (a), the same is true for $\mathbf{N} + \mathbf{N}^T$.

(b: Alternate) Since the eigenvalues of $\mathbf{N}+\mathbf{N}^T$ are $\lambda_j = 2\cos(j\pi/n+1)$ you can argue that $\mathbf{N}+\mathbf{N}^T$ has a zero eigenvalue (and hence is singular) if and only if n is odd by showing that there exists an integer α such that $j\pi/n+1 = \alpha\pi/2$ for some $1 \le j \le n$ if and only if n is odd.

(c) Since a determinant is the product of eigenvalues (recall (7.1.8), p. 494), $\det\left(\mathbf{N} - \mathbf{N}^T\right)/\det\left(\mathbf{N} + \mathbf{N}^T\right) = (\mathrm{i}\lambda_1 \cdots \mathrm{i}\lambda_n)/(\lambda_1 \cdots \lambda_n) = \mathrm{i}^n = (-1)^{n/2}$.

7.2.20. The eigenvalues are $\{2, 0, 2, 0\}$. The columns of $\mathbf{F}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \end{pmatrix}$ are corresponding eigenvectors.

7.2.21. $\mathbf{Ax} = \lambda\mathbf{x} \implies \mathbf{y}^*\mathbf{Ax} = \lambda\mathbf{y}^*\mathbf{x}$ and $\mathbf{y}^*\mathbf{A} = \mu\mathbf{y}^* \implies \mathbf{y}^*\mathbf{Ax} = \mu\mathbf{y}^*\mathbf{x}$. Therefore, $\lambda\mathbf{y}^*\mathbf{x} = \mu\mathbf{y}^*\mathbf{x} \implies (\lambda - \mu)\mathbf{y}^*\mathbf{x} = 0 \implies \mathbf{y}^*\mathbf{x} = 0$ when $\lambda \neq \mu$.

7.2.22. (a) Suppose $\mathbf{P}$ is a nonsingular matrix such that $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}$ is diagonal, and suppose that λ is the kth diagonal entry in $\mathbf{D}$. If $\mathbf{x}$ and $\mathbf{y}^*$ are the kth column and k^{th} row in $\mathbf{P}$ and $\mathbf{P}^{-1}$, respectively, then $\mathbf{x}$ and $\mathbf{y}^*$ must be right-hand and left-hand eigenvectors associated with λ such that $\mathbf{y}^*\mathbf{x} = 1$.

(b) Consider $\mathbf{A} = \mathbf{I}$ with $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$ for $i \neq j$.

(c) Consider $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

7.2.23. (a) Suppose not—i.e., suppose $\mathbf{y}^*\mathbf{x} = 0$. Then

$$\mathbf{x} \perp span\,(\mathbf{y}) = N\left(\mathbf{A} - \lambda\mathbf{I}\right)^* \implies \mathbf{x} \in N\left(\mathbf{A} - \lambda\mathbf{I}\right)^{*\perp} = R\left(\mathbf{A} - \lambda\mathbf{I}\right).$$

Also, $\mathbf{x} \in N(\mathbf{A} - \lambda\mathbf{I})$, so $\mathbf{x} \in R(\mathbf{A} - \lambda\mathbf{I}) \cap N(\mathbf{A} - \lambda\mathbf{I})$. However, because λ is a simple eigenvalue, the the core-nilpotent decomposition on p. 397 insures that $\mathbf{A} - \lambda\mathbf{I}$ is similar to a matrix of the form $\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & 0_{1\times 1} \end{pmatrix}$, and this implies that $R(\mathbf{A} - \lambda\mathbf{I}) \cap N(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$ (Exercise 5.10.12, p. 402), which is a contradiction. Thus $\mathbf{y}^*\mathbf{x} \neq 0$.

(b) Consider $\mathbf{A} = \mathbf{I}$ with $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$ for $i \neq j$.

7.2.24. Let $\mathcal{B}_i$ be a basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$, and suppose $\mathbf{A}$ is diagonalizable. Since $\textit{geo mult}_{\mathbf{A}}(\lambda_i) = \textit{alg mult}_{\mathbf{A}}(\lambda_i)$ for each i, (7.2.4) implies $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ is a set of n independent vectors—i.e., $\mathcal{B}$ is a basis for $\Re^n$. Exercise 5.9.14 now guarantees that $\Re^n = N(\mathbf{A} - \lambda_1\mathbf{I}) \oplus N(\mathbf{A} - \lambda_2\mathbf{I}) \oplus \cdots \oplus N(\mathbf{A} - \lambda_k\mathbf{I})$. Conversely, if this equation holds, then Exercise 5.9.14 says $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$ is a basis for $\Re^n$, and hence $\mathbf{A}$ is diagonalizable because $\mathcal{B}$ is a complete independent set of eigenvectors.

7.2.25. Proceed inductively just as in the development of Schur's triangularization theorem. If the first eigenvalue λ is real, the reduction is exactly the same as described on p. 508 (with everything being real). If λ is complex, then $(\lambda, \mathbf{x})$ and $(\overline{\lambda}, \overline{\mathbf{x}})$ are both eigenpairs for $\mathbf{A}$, and, by (7.2.3), $\{\mathbf{x}, \overline{\mathbf{x}}\}$ is linearly independent. Consequently, if $\mathbf{x} = \mathbf{u} + \mathrm{i}\mathbf{v}$, with $\mathbf{u}, \mathbf{v} \in \Re^{n\times 1}$, then $\{\mathbf{u}, \mathbf{v}\}$ is linearly independent—otherwise, $\mathbf{u} = \xi\mathbf{v}$ implies $\mathbf{x} = (1 + \mathrm{i}\xi)\mathbf{u}$ and $\overline{\mathbf{x}} = (1 - \mathrm{i}\xi)\mathbf{u}$, which is impossible. Let $\lambda = \alpha + \mathrm{i}\beta$, $\alpha, \beta \in \Re$, and observe that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ implies $\mathbf{A}\mathbf{u} = \alpha\mathbf{u} - \beta\mathbf{v}$ and $\mathbf{A}\mathbf{v} = \beta\mathbf{u} + \alpha\mathbf{v}$, so $\mathbf{A}\mathbf{W} = \mathbf{W}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}$, where $\mathbf{W} = \begin{bmatrix} \mathbf{u} \,|\, \mathbf{v} \end{bmatrix}$. Let $\mathbf{W} = \mathbf{Q}_{n\times 2}\mathbf{R}_{2\times 2}$ be a rectangular QR factorization (p. 311), and let $\mathbf{B} = \mathbf{R}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}\mathbf{R}^{-1}$ so that $\sigma(\mathbf{B}) = \sigma\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} = \{\lambda, \overline{\lambda}\}$, and

$$\mathbf{A}\mathbf{W} = \mathbf{A}\mathbf{Q}\mathbf{R} = \mathbf{Q}\mathbf{R}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \implies \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{R}\begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}\mathbf{R}^{-1} = \mathbf{B}.$$

If $\mathbf{X}_{n\times n-2}$ is chosen so that $\mathbf{P} = \begin{bmatrix} \mathbf{Q} \,|\, \mathbf{X} \end{bmatrix}$ is an orthogonal matrix (i.e., the columns of $\mathbf{X}$ complete the two columns of $\mathbf{Q}$ to an orthonormal basis for $\Re^n$), then $\mathbf{X}^T\mathbf{A}\mathbf{Q} = \mathbf{X}^T\mathbf{Q}\mathbf{B} = \mathbf{0}$, and

$$\mathbf{P}^T\mathbf{A}\mathbf{P} = \begin{pmatrix} \mathbf{Q}^T\mathbf{A}\mathbf{Q} & \mathbf{Q}^T\mathbf{A}\mathbf{X} \\ \mathbf{X}^T\mathbf{A}\mathbf{Q} & \mathbf{X}^T\mathbf{A}\mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{B} & \mathbf{Q}^T\mathbf{A}\mathbf{X} \\ \mathbf{0} & \mathbf{X}^T\mathbf{A}\mathbf{X} \end{pmatrix}.$$

Now repeat the argument on the $n-2 \times n-2$ matrix $\mathbf{X}^T\mathbf{A}\mathbf{X}$. Continuing in this manner produces the desired conclusion.

7.2.26. Let the columns $\mathbf{R}_{n\times r}$ be linearly independent eigenvectors corresponding to the real eigenvalues ρ_j, and let $\{\mathbf{x}_1, \overline{\mathbf{x}}_1, \mathbf{x}_2, \overline{\mathbf{x}}_2, \ldots, \mathbf{x}_t, \overline{\mathbf{x}}_t\}$ be a set of linearly independent eigenvectors associated with $\{\lambda_1, \overline{\lambda}_1, \lambda_2, \overline{\lambda}_2, \ldots, \lambda_t, \overline{\lambda}_t\}$ so that the matrix $\mathbf{Q} = \begin{bmatrix} \mathbf{R} \,|\, \mathbf{x}_1 \,|\, \overline{\mathbf{x}}_1 \,|\, \cdots \,|\, \mathbf{x}_t \,|\, \overline{\mathbf{x}}_t \end{bmatrix}$ is nonsingular. Write $\mathbf{x}_j = \mathbf{u}_j + \mathrm{i}\mathbf{v}_j$ for

$\mathbf{u}_j, \mathbf{v}_j \in \Re^{n\times 1}$ and $\lambda_j = \alpha_j + \mathrm{i}\beta_j$ for $\alpha, \beta \in \Re$, and let $\mathbf{P}$ be the real matrix $\mathbf{P} = \left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\, \mathbf{v}_1 \,|\, \mathbf{u}_2 \,|\, \mathbf{v}_2 \,|\, \cdots \,|\, \mathbf{u}_t \,|\, \mathbf{v}_t\right]$. This matrix is nonsingular because Exercise 6.1.14 can be used to show that $\det(\mathbf{P}) = 2t(-\mathrm{i})^t \det(\mathbf{Q})$. For example, if $t = 1$, then $\mathbf{P} = \left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\, \mathbf{v}_1\right]$ and

$$\begin{aligned}
\det(\mathbf{Q}) &= \det\left[\mathbf{R} \,|\, \mathbf{x}_1 \,|\, \overline{\mathbf{x}}_1\right] = \det\left[\mathbf{R} \,|\, \mathbf{u}_1 + \mathrm{i}\mathbf{v}_1 \,|\, \mathbf{u}_1 - \mathrm{i}\mathbf{v}_1\right] \\
&= \det\left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\, \mathbf{u}_1\right] + \det\left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\, -\mathrm{i}\mathbf{v}_1\right] \\
&\quad + \det\left[\mathbf{R} \,|\, \mathrm{i}\mathbf{v}_1 \,|\, \mathbf{u}_1\right] + \det\left[\mathbf{R} \,|\, \mathrm{i}\mathbf{v}_1 \,|\, \mathrm{i}\mathbf{v}_1\right] \\
&= -\mathrm{i}\det\left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\mathbf{v}_1\right] + \mathrm{i}\det\left[\mathbf{R} \,|\, \mathbf{v}_1 \,|\, \mathbf{u}_1\right] \\
&= -\mathrm{i}\det\left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\mathbf{v}_1\right] - \mathrm{i}\det\left[\mathbf{R} \,|\, \mathbf{u}_1 \,|\mathbf{v}_1\right] = 2(-\mathrm{i})\det(\mathbf{P}).
\end{aligned}$$

Induction can now be used. The equations $\mathbf{A}(\mathbf{u}_j + \mathrm{i}\mathbf{v}_j) = (\alpha_j + \mathrm{i}\beta_j)(\mathbf{u}_j + \mathrm{i}\mathbf{v}_j)$ yield $\mathbf{A}\mathbf{u}_j = \alpha_j\mathbf{u}_j - \beta_j\mathbf{v}_j$ and $\mathbf{A}\mathbf{v}_j = \beta_j\mathbf{u}_j + \alpha_j\mathbf{v}_j$. Couple these with the fact that $\mathbf{AR} = \mathbf{RD}$ to conclude that

$$\mathbf{AP} = \left[\mathbf{RD} \,|\, \cdots \,|\, \alpha_j\mathbf{u}_j - \beta_j\mathbf{v}_j \,|\, \beta_j\mathbf{u}_j + \alpha_j\mathbf{v}_j \,|\, \cdots\right] = \mathbf{P}\begin{pmatrix} \mathbf{D} & 0 & \cdots & 0 \\ 0 & \mathbf{B}_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{B}_t \end{pmatrix},$$

where

$$\mathbf{D} = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho_r \end{pmatrix} \quad \text{and} \quad \mathbf{B}_j = \begin{pmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{pmatrix}.$$

7.2.27. Schur's triangularization theorem says $\mathbf{U}^*\mathbf{AU} = \mathbf{T}$ where $\mathbf{U}$ is unitary and $\mathbf{T}$ is upper triangular. Setting $\mathbf{x} = \mathbf{U}\mathbf{e}_i$ in $\mathbf{x}^*\mathbf{Ax} = 0$ yields that $t_{ii} = 0$ for each i, so $t_{ij} = 0$ for all $i \geq j$. Now set $\mathbf{x} = \mathbf{U}(\mathbf{e}_i + \mathbf{e}_j)$ with $i < j$ in $\mathbf{x}^*\mathbf{Ax} = 0$ to conclude that $t_{ij} = 0$ whenever $i < j$. Consequently, $\mathbf{T} = \mathbf{0}$, and thus $\mathbf{A} = \mathbf{0}$. To see that $\mathbf{x}^T\mathbf{Ax} = 0 \;\forall\; \mathbf{x} \in \Re^{n\times 1} \not\Rightarrow \mathbf{A} = \mathbf{0}$, consider $\mathbf{A} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

Solutions for exercises in section 7. 3

7.3.1. $\cos\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The characteristic equation for $\mathbf{A}$ is $\lambda^2 + \pi\lambda = 0$, so the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 0$ and $\lambda_2 = -\pi$. Note that $\mathbf{A}$ is diagonalizable because no eigenvalue is repeated. Associated eigenvectors are computed in the usual way to be

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

so

$$\mathbf{P} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

Thus

$$\cos \mathbf{A} = \mathbf{P}\begin{pmatrix} \cos(0) & 0 \\ 0 & \cos(-\pi) \end{pmatrix}\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

7.3.2. From Example 7.3.3, the eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = -(\alpha+\beta)$, and associated eigenvectors are computed in the usual way to be

$$\mathbf{x}_1 = \begin{pmatrix} \beta/\alpha \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

so

$$\mathbf{P} = \begin{pmatrix} \beta/\alpha & -1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{P}^{-1} = \frac{1}{1+\beta/\alpha}\begin{pmatrix} 1 & 1 \\ -1 & \beta/\alpha \end{pmatrix}.$$

Thus

$$\mathbf{P}\begin{pmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{pmatrix}\mathbf{P}^{-1} = \frac{\alpha}{\alpha+\beta}\begin{pmatrix} \beta/\alpha & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & e^{-(\alpha+\beta)t} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & \beta/\alpha \end{pmatrix}$$
$$= \frac{1}{\alpha+\beta}\left[\begin{pmatrix} \beta & \beta \\ \alpha & \alpha \end{pmatrix} + e^{-(\alpha+\beta)t}\begin{pmatrix} \alpha & -\beta \\ -\alpha & \beta \end{pmatrix}\right]$$
$$= e^{\lambda_1 t}\mathbf{G}_1 + e^{\lambda_2 t}\mathbf{G}_2.$$

7.3.3. *Solution* 1: If $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then

$$\sin^2 \mathbf{A} = \mathbf{P}\left(\sin^2 \mathbf{D}\right)\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \sin^2\lambda_1 & 0 & \cdots & 0 \\ 0 & \sin^2\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sin^2\lambda_n \end{pmatrix}\mathbf{P}^{-1}.$$

Similarly for $\cos^2\mathbf{A}$, so $\sin^2\mathbf{A} + \cos^2\mathbf{A} = \mathbf{P}\left(\sin^2\mathbf{D} + \cos^2\mathbf{D}\right)\mathbf{P}^{-1} = \mathbf{P}\mathbf{I}\mathbf{P}^{-1} = \mathbf{I}$.

Solution 2: If $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, use the spectral representation (7.3.6) to write $\sin^2\mathbf{A} = \sum_{i=1}^k (\sin^2\lambda_i)\mathbf{G}_i$ and $\cos^2\mathbf{A} = \sum_{i=1}^k (\cos^2\lambda_i)\mathbf{G}_i$, so that $\sin^2\mathbf{A} + \cos^2\mathbf{A} = \sum_{i=1}^k (\sin^2\lambda_i + \cos^2\lambda_i)\mathbf{G}_i = \sum_{i=1}^k \mathbf{G}_i = \mathbf{I}$.

7.3.4. The infinite series representation of $e^{\mathbf{A}}$ readily yields this.

7.3.5. (a) Eigenvalues are invariant under a similarity transformation, so the eigenvalues of $f(\mathbf{A}) = \mathbf{P}f(\mathbf{D})\mathbf{P}^{-1}$ are the eigenvalues of $f(\mathbf{D})$, which are given by $\{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n)\}$.

(b) If $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A}$, then $(\mathbf{A} - z_0\mathbf{I})^n\mathbf{x} = (\lambda - z_0)^n\mathbf{x}$ implies that $(f(\lambda), \mathbf{x})$ is an eigenpair for $f(\mathbf{A})$.

7.3.6. If $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are the eigenvalues of $\mathbf{A}_{n\times n}$, then $\{e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n}\}$ are the eigenvalues of $e^{\mathbf{A}}$ by the spectral mapping property from Exercise 7.3.5. The trace is the sum of the eigenvalues, and the determinant is the product of the eigenvalues (p. 494), so $\det\left(e^{\mathbf{A}}\right) = e^{\lambda_1}e^{\lambda_2}\cdots e^{\lambda_n} = e^{\lambda_1+\lambda_2+\cdots+\lambda_n} = e^{trace(\mathbf{A})}$.

7.3.7. The Cayley–Hamilton theorem says that each $\mathbf{A}_{m\times m}$ satisfies its own characteristic equation, $0 = \det(\mathbf{A} - \lambda\mathbf{I}) = \lambda^m + c_1\lambda^{m-1} + c_2\lambda^{m-2} + \cdots + c_{m-1}\lambda + c_m$, so $\mathbf{A}^m = -c_1\mathbf{A}^{m-1} - \cdots - c_{m-1}\mathbf{A} - c_m\mathbf{I}$. Consequently, $\mathbf{A}^m$ and every higher power of $\mathbf{A}$ is a polynomial in $\mathbf{A}$ of degree at most $m-1$, and thus any expression involving powers of $\mathbf{A}$ can always be reduced to an expression involving at most $\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{m-1}$.

7.3.8. When $\mathbf{A}$ is diagonalizable, (7.3.11) insures $f(\mathbf{A}) = p(\mathbf{A})$ is a polynomial in $\mathbf{A}$, and $\mathbf{A}p(\mathbf{A}) = p(\mathbf{A})\mathbf{A}$. If $f(\mathbf{A})$ is defined by the series (7.3.7) in the nondiagonalizable case, then, by Exercise 7.3.7, it's still true that $f(\mathbf{A}) = p(\mathbf{A})$ is a polynomial in $\mathbf{A}$, and thus $\mathbf{A}f(\mathbf{A}) = f(\mathbf{A})\mathbf{A}$ holds in the nondiagonalizable case also.

7.3.9. If $\mathbf{A}$ and $\mathbf{B}$ are diagonalizable with $\mathbf{AB} = \mathbf{BA}$, Exercise 7.2.16 insures $\mathbf{A}$ and $\mathbf{B}$ can be simultaneously diagonalized. If $\mathbf{P}^{-1}\mathbf{AP} = \mathbf{D}_A = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\mathbf{P}^{-1}\mathbf{BP} = \mathbf{D}_B = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n)$, then $\mathbf{A} + \mathbf{B} = \mathbf{P}(\mathbf{D}_A + \mathbf{D}_B)\mathbf{P}^{-1}$, so

$$
\begin{aligned}
e^{\mathbf{A}+\mathbf{B}} &= \mathbf{P}\left(e^{\mathbf{D}_A+\mathbf{D}_B}\right)\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} e^{\lambda_1+\mu_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2+\mu_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n+\mu_n} \end{pmatrix}\mathbf{P}^{-1} \\
&= \mathbf{P}\begin{pmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n} \end{pmatrix}\mathbf{P}^{-1}\mathbf{P}\begin{pmatrix} e^{\mu_1} & 0 & \cdots & 0 \\ 0 & e^{\mu_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\mu_n} \end{pmatrix}\mathbf{P}^{-1} \\
&= e^{\mathbf{A}}e^{\mathbf{B}}.
\end{aligned}
$$

In general, the same brute force multiplication of scalar series that yields

$$
e^{x+y} = \sum_{n=0}^{\infty}\frac{(x+y)^n}{n!} = \left(\sum_{n=0}^{\infty}\frac{x^n}{n!}\right)\left(\sum_{n=0}^{\infty}\frac{y^n}{n!}\right) = e^x e^y
$$

holds for matrix series when $\mathbf{AB} = \mathbf{BA}$, but this is quite messy. A more elegant approach is to set $\mathbf{F}(t) = e^{\mathbf{A}t+\mathbf{B}t} - e^{\mathbf{A}t}e^{\mathbf{B}t}$ and note that $\mathbf{F}'(t) = \mathbf{0}$ for all t when $\mathbf{AB} = \mathbf{BA}$, so $\mathbf{F}(t)$ must be a constant matrix for all t. Since $\mathbf{F}(0) = \mathbf{0}$, it follows that $e^{(\mathbf{A}+\mathbf{B})t} = e^{\mathbf{A}t}e^{\mathbf{B}t}$ for all t. To see that $e^{\mathbf{A}+\mathbf{B}}$, $e^{\mathbf{A}}e^{\mathbf{B}}$, and $e^{\mathbf{B}}e^{\mathbf{A}}$ can be different when $\mathbf{AB} \neq \mathbf{BA}$, consider $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

7.3.10. The infinite series representation of $e^{\mathbf{A}}$ shows that if $\mathbf{A}$ is skew symmetric, then $\left(e^{\mathbf{A}}\right)^T = e^{\mathbf{A}^T} = \mathbf{e}^{-\mathbf{A}}$, and hence $e^{\mathbf{A}}\left(e^{\mathbf{A}}\right)^T = e^{\mathbf{A}-\mathbf{A}} = e^{\mathbf{0}} = \mathbf{I}$.

7.3.11. (a) Draw a transition diagram similar to that in Figure 7.3.1 with North and South replaced by ON and OFF, respectively. Let x_k be the fraction of switches in the ON state and let y_k be the fraction of switches in the OFF state after k clock cycles have elapsed. According to the given information,

$$\begin{aligned} x_k &= x_{k-1}(.1) + y_{k-1}(.3) \\ y_k &= x_{k-1}(.9) + y_{k-1}(.7) \end{aligned}$$

so that $\mathbf{p}_{k+1}^T = \mathbf{p}_k^T\mathbf{T}$, where $\mathbf{p}_k^T = (\,x_k \quad y_k\,)$ and $\mathbf{T} = \begin{pmatrix} .1 & .9 \\ .3 & .7 \end{pmatrix}$. Compute $\sigma(\mathbf{T}) = \{1, -1/5\}$, and use the methods of Example 7.3.4 to determine the steady-state (or limiting) distribution as

$$\begin{aligned} \mathbf{p}_\infty^T &= \lim_{k\to\infty} \mathbf{p}_k^T = \lim_{k\to\infty} \mathbf{p}_0^T\mathbf{T}^k = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{T}^k = (\,x_0 \quad y_0\,)\begin{pmatrix} 1/4 & 3/4 \\ 1/4 & 3/4 \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{x_0+y_0}{4} & \dfrac{3(x_0+y_0)}{4} \end{pmatrix} = (\,1/4 \quad 3/4\,). \end{aligned}$$

Alternately, (7.3.15) can be used with $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mathbf{y}_1 = (\,1 \quad 3\,)$ to obtain

$$\mathbf{p}_\infty^T = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{T}^k = \mathbf{p}_0^T \lim_{k\to\infty} \mathbf{G}_1 = \frac{(\mathbf{p}_0^T\mathbf{x}_1)\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1} = \frac{\mathbf{y}_1^T}{\mathbf{y}_1^T\mathbf{x}_1} = (\,1/4 \quad 3/4\,).$$

(b) Computing a few powers of $\mathbf{T}$ reveals that

$$\mathbf{T}^2 = \begin{pmatrix} .280 & .720 \\ .240 & .760 \end{pmatrix}, \quad \mathbf{T}^3 = \begin{pmatrix} .244 & .756 \\ .252 & .748 \end{pmatrix},$$

$$\mathbf{T}^4 = \begin{pmatrix} .251 & .749 \\ .250 & .750 \end{pmatrix}, \quad \mathbf{T}^5 = \begin{pmatrix} .250 & .750 \\ .250 & .750 \end{pmatrix},$$

so, for practical purposes, the device can be considered to be in equilibrium after about 5 clock cycles, regardless of the initial configuration.

7.3.12. Let $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_k|$, and assume $\lambda_1 \neq 0$; otherwise $\mathbf{A} = \mathbf{0}$ and there is nothing to prove. Set

$$\begin{aligned} \nu_n &= \frac{\|\mathbf{A}^n\|}{|\lambda_1^n|} = \frac{\|\lambda_1^n\mathbf{G}_1 + \lambda_2^n\mathbf{G}_2 + \cdots + \lambda_k^n\mathbf{G}_k\|}{|\lambda_1^n|} = \left\| \frac{\lambda_1^n\mathbf{G}_1 + \lambda_2^n\mathbf{G}_2 + \cdots + \lambda_k^n\mathbf{G}_k}{\lambda_1^n} \right\| \\ &= \left\| \mathbf{G}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^n \mathbf{G}_2 + \cdots + \left(\frac{\lambda_k}{\lambda_1}\right)^n \mathbf{G}_k \right\| \quad \text{and let} \quad \nu = \sum_{i=1}^k \|\mathbf{G}_i\|. \end{aligned}$$

Observe that $1 \le \nu_n \le \nu$ for every positive integer n—the first inequality follows because $\lambda_1^n \in \sigma(\mathbf{A}^n)$ implies $|\lambda_1^n| \le \|\mathbf{A}^n\|$ by (7.1.12) on p. 497, and the second is the result of the triangle inequality. Consequently,

$$1^{1/n} \le \nu_n^{1/n} \le \nu^{1/n} \Longrightarrow 1 \le \lim_{n\to\infty} \nu_n^{1/n} \le 1 \Longrightarrow 1 = \lim_{n\to\infty} \nu_n^{1/n} = \lim_{n\to\infty} \frac{\|\mathbf{A}^n\|^{1/n}}{|\lambda_1|}.$$

7.3.13. The dominant eigenvalue is $\lambda_1 = 4$, and all corresponding eigenvectors are multiples of $(-1, 0, 1)^T$.

7.3.15. Consider

$$\mathbf{x}_n = \begin{pmatrix} 1 - 1/n \\ -1 \end{pmatrix} \to \mathbf{x} = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

but $m(\mathbf{x}_n) = -1$ for all $n = 1, 2, \ldots,$ and $m(\mathbf{x}) = 1$, so $m(\mathbf{x}_n) \not\to m(\mathbf{x})$. Nevertheless, if $\lim_{n\to\infty} \mathbf{x}_n \ne \mathbf{0}$, then $\lim_{n\to\infty} m(\mathbf{x}_n) \ne 0$ because the function $\tilde{m}(\mathbf{v}) = |m(\mathbf{v})| = \|\mathbf{v}\|_\infty$ is continuous.

7.3.16. (a) The "vanilla" QR iteration fails to converge.

(b) $\mathbf{H} - \mathbf{I} = \mathbf{QR} = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 3 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and $\mathbf{RQ} + \mathbf{I} = \begin{pmatrix} -2 & 1 & 1 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Solutions for exercises in section 7. 4

7.4.1. The unique solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, is

$$\begin{aligned}
\mathbf{u} = \mathrm{e}^{\mathbf{A}t}\mathbf{c} &= \mathbf{P} \begin{pmatrix} \mathrm{e}^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & \mathrm{e}^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{e}^{\lambda_n t} \end{pmatrix} \mathbf{P}^{-1}\mathbf{c} \\
&= [\mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \cdots \,|\, \mathbf{x}_n] \begin{pmatrix} \mathrm{e}^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & \mathrm{e}^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{e}^{\lambda_n t} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} \\
&= \xi_1 \mathrm{e}^{\lambda_1 t}\mathbf{x}_1 + \xi_2 \mathrm{e}^{\lambda_2 t}\mathbf{x}_2 + \cdots + \xi_n \mathrm{e}^{\lambda_n t}\mathbf{x}_n.
\end{aligned}$$

7.4.2. (a) All eigenvalues in $\sigma(\mathbf{A}) = \{-1, -3\}$ are negative, so the system is stable.

(b) All eigenvalues in $\sigma(\mathbf{A}) = \{1, 3\}$ are positive, so the system is unstable.

(c) $\sigma(\mathbf{A}) = \{\pm \mathrm{i}\}$, so the system is semistable. If $\mathbf{c} \ne \mathbf{0}$, then the components in $\mathbf{u}(t)$ will oscillate indefinitely.

7.4.3. (a) If $u_k(t)$ denotes the number in population k at time t, then

$$\begin{aligned}
u_1' &= 2u_1 - u_2, \qquad & u_1(0) &= 100, \\
u_2' &= -u_1 + 2u_2, \qquad & u_2(0) &= 200,
\end{aligned}$$

or $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} 100 \\ 200 \end{pmatrix}$. The characteristic equation for $\mathbf{A}$ is $p(\lambda) = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3) = 0$, so the eigenvalues for $\mathbf{A}$ are $\lambda_1 = 1$ and $\lambda_2 = 3$. We know from (7.4.7) that

$$\mathbf{u}(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 \quad (\text{where } \mathbf{v}_i = \mathbf{G}_i\mathbf{c}\,)$$

is the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. The spectral theorem on p. 517 implies $\mathbf{A} - \lambda_2\mathbf{I} = (\lambda_1 - \lambda_2)\mathbf{G}_1$ and $\mathbf{I} = \mathbf{G}_1 + \mathbf{G}_2$, so $(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c} = (\lambda_1 - \lambda_2)\mathbf{v}_1$ and $\mathbf{c} = \mathbf{v}_1 + \mathbf{v}_2$, and consequently

$$\mathbf{v}_1 = \frac{(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c}}{(\lambda_1 - \lambda_2)} = \begin{pmatrix} 150 \\ 150 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \mathbf{c} - \mathbf{v}_1 = \begin{pmatrix} -50 \\ 50 \end{pmatrix},$$

so

$$u_1(t) = 150\mathrm{e}^t - 50\mathrm{e}^{3t} \quad \text{and} \quad u_2(t) = 150\mathrm{e}^t + 50\mathrm{e}^{3t}.$$

(b) As $t \to \infty$, $u_1(t) \to -\infty$ and $u_2(t) \to +\infty$. But a population can't become negative, so species I is destined to become extinct, and this occurs at the value of t for which $u_1(t) = 0$—i.e., when

$$\mathrm{e}^t\left(\mathrm{e}^{2t} - 3\right) = 0 \implies \mathrm{e}^{2t} = 3 \implies t = \frac{\ln 3}{2}.$$

7.4.4. If $u_k(t)$ denotes the number in population k at time t, then the hypothesis says

$$\begin{aligned} u_1' &= -u_1 + u_2, \qquad & u_1(0) &= 200, \\ u_2' &= u_1 - 2u_2, \qquad & u_2(0) &= 400, \end{aligned}$$

or $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$, where $\mathbf{A} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\mathbf{c} = \begin{pmatrix} 200 \\ 400 \end{pmatrix}$. The characteristic equation for $\mathbf{A}$ is $p(\lambda) = \lambda^2 + 2\lambda = \lambda(\lambda + 2) = 0$, so the eigenvalues for $\mathbf{A}$ are $\lambda_1 = 0$ and $\lambda_2 = -2$. We know from (7.4.7) that

$$\mathbf{u}(t) = \mathrm{e}^{\lambda_1 t}\mathbf{v}_1 + \mathrm{e}^{\lambda_2 t}\mathbf{v}_2 \quad (\text{where } \mathbf{v}_i = \mathbf{G}_i\mathbf{c}\,)$$

is the solution to $\mathbf{u}' = \mathbf{A}\mathbf{u}$, $\mathbf{u}(0) = \mathbf{c}$. The spectral theorem on p. 517 implies $\mathbf{A} - \lambda_2\mathbf{I} = (\lambda_1 - \lambda_2)\mathbf{G}_1$ and $\mathbf{I} = \mathbf{G}_1 + \mathbf{G}_2$, so $(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c} = (\lambda_1 - \lambda_2)\mathbf{v}_1$ and $\mathbf{c} = \mathbf{v}_1 + \mathbf{v}_2$, and consequently

$$\mathbf{v}_1 = \frac{(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{c}}{(\lambda_1 - \lambda_2)} = \begin{pmatrix} 300 \\ 300 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \mathbf{c} - \mathbf{v}_1 = \begin{pmatrix} -100 \\ 100 \end{pmatrix},$$

so

$$u_1(t) = 300 - 100\mathrm{e}^{-2t} \quad \text{and} \quad u_2(t) = 300 + 100\mathrm{e}^{-2t}.$$

As $t \to \infty$, $u_1(t) \to 300$ and $u_2(t) \to 300$, so both populations will stabilize at 300.

Solutions for exercises in section 7. 5

7.5.1. Yes, because $\mathbf{A}^*\mathbf{A} = \mathbf{A}\mathbf{A}^* = \begin{pmatrix} 30 & 6-6\,\mathrm{i} \\ 6+6\,\mathrm{i} & 24 \end{pmatrix}$.

7.5.2. Real skew-symmetric and orthogonal matrices are examples.

7.5.3. We already know from (7.5.3) that real-symmetric matrices are normal and have real eigenvalues, so only the converse needs to be proven. If $\mathbf{A}$ is real and normal with real eigenvalues, then there is a complete orthonormal set of real eigenvectors, so using them as columns in $\mathbf{P} \in \Re^{n\times n}$ results in an orthogonal matrix such that $\mathbf{P}^T\mathbf{A}\mathbf{P} = \mathbf{D}$ is diagonal or, equivalently, $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T$, and thus $\mathbf{A} = \mathbf{A}^T$.

7.5.4. If $(\lambda, \mathbf{x})$ is an eigenpair for $\mathbf{A} = -\mathbf{A}^*$ then $\mathbf{x}^*\mathbf{x} \neq 0$, and $\lambda\mathbf{x} = \mathbf{A}\mathbf{x}$ implies $\overline{\lambda}\mathbf{x}^* = \mathbf{x}^*\mathbf{A}^*$, so

$$\mathbf{x}^*\mathbf{x}(\lambda + \overline{\lambda}) = \mathbf{x}^*(\lambda + \overline{\lambda})\mathbf{x} = \mathbf{x}^*\mathbf{A}\mathbf{x} + \mathbf{x}^*\mathbf{A}^*\mathbf{x} = 0 \implies \lambda = -\overline{\lambda} \implies \Re\mathrm{e}(\lambda) = 0.$$

7.5.5. If $\mathbf{A}$ is skew hermitian (real skew symmetric), then $\mathbf{A}$ is normal, and hence $\mathbf{A}$ is unitarily (orthogonally) similar to a diagonal matrix—say $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$. Moreover, the eigenvalues λ_j in $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are pure imaginary numbers (Exercise 7.5.4). Since $f(z) = (1-z)(1+z)^{-1}$ maps the imaginary axis in the complex plane to points on the unit circle, each $f(\lambda_j)$ is on the unit circle, so there is some θ_j such that $f(\lambda_j) = \mathrm{e}^{\mathrm{i}\theta_j} = \cos\theta_j + \mathrm{i}\sin\theta_j$. Consequently,

$$f(\mathbf{A}) = \mathbf{U}\begin{pmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_n) \end{pmatrix}\mathbf{U}^* = \mathbf{U}\begin{pmatrix} \mathrm{e}^{\mathrm{i}\theta_1} & 0 & \cdots & 0 \\ 0 & \mathrm{e}^{\mathrm{i}\theta_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{e}^{\mathrm{i}\theta_n} \end{pmatrix}\mathbf{U}^*$$

together with $\mathrm{e}^{\mathrm{i}\theta_j}\overline{\mathrm{e}^{\mathrm{i}\theta_j}} = \mathrm{e}^{\mathrm{i}\theta_j}\mathrm{e}^{-\mathrm{i}\theta_j} = 1$ yields $f(\mathbf{A})^*f(\mathbf{A}) = \mathbf{I}$. **Note:** The fact that $(\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A})^{-1} = (\mathbf{I}+\mathbf{A})^{-1}(\mathbf{I}-\mathbf{A})$ follows from Exercise 7.3.8. See the solution to Exercise 5.6.6 for an alternate approach.

7.5.6. Consider the identity matrix—every nonzero vector is an eigenvector, so not every complete independent set of eigenvectors needs to be orthonormal. Given a complete independent set of eigenvectors for a normal $\mathbf{A}$ with $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$, use the Gram–Schmidt procedure to form an orthonormal basis for $N(\mathbf{A} - \lambda_i\mathbf{I})$ for each i. Since $N(\mathbf{A} - \lambda_i\mathbf{I}) \perp N(\mathbf{A} - \lambda_j\mathbf{I})$ for $\lambda_i \neq \lambda_j$ (by (7.5.2)), the union of these orthonormal bases will be a complete orthonormal set of eigenvectors for $\mathbf{A}$.

7.5.7. Consider $\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

7.5.8. Suppose $\mathbf{T}_{n\times n}$ is an upper-triangular matrix such that $\mathbf{T}^*\mathbf{T} = \mathbf{T}\mathbf{T}^*$. The (1,1)-entry of $\mathbf{T}^*\mathbf{T}$ is $|t_{11}|^2$, and the (1,1)-entry of $\mathbf{T}\mathbf{T}^*$ is $\sum_{k=1}^n |t_{1k}|^2$. Equating

these implies $t_{12} = t_{13} = \cdots = t_{1n} = 0$. Now use this and compare the (2,2)-entries to get $t_{23} = t_{24} = \cdots = t_{2n} = 0$. Repeating this argument for each row produces the conclusion that $\mathbf{T}$ must be diagonal. Conversely, if $\mathbf{T}$ is diagonal, then $\mathbf{T}$ is normal because $\mathbf{T}^*\mathbf{T} = \operatorname{diag}(|t_{11}|^2 \cdots |t_{nn}|^2) = \mathbf{T}\mathbf{T}^*$.

7.5.9. Schur's triangularization theorem on p. 508 says every square matrix is unitarily similar to an upper-triangular matrix—say $\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{T}$. If $\mathbf{A}$ is normal, then so is $\mathbf{T}$. Exercise 7.5.8 therefore insures that $\mathbf{T}$ must be diagonal. Conversely, if $\mathbf{T}$ is diagonal, then it is normal, and thus so is $\mathbf{A}$.

7.5.10. If $\mathbf{A}$ is normal, so is $\mathbf{A} - \lambda\mathbf{I}$. Consequently, $\mathbf{A} - \lambda\mathbf{I}$ is RPN, and hence $N(\mathbf{A} - \lambda\mathbf{I}) = N(\mathbf{A} - \lambda\mathbf{I})^*$ (p. 408), so $(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0} \Longleftrightarrow (\mathbf{A}^* - \overline{\lambda}\mathbf{I})\mathbf{x} = \mathbf{0}$.

7.5.11. Just as in the proof of the min-max part, it suffices to prove

$$\lambda_i = \max_{\dim \mathcal{V} = i} \; \min_{\substack{\mathbf{y} \in \mathcal{V} \\ \|\mathbf{y}\|_2 = 1}} \mathbf{y}^*\mathbf{D}\mathbf{y}.$$

For each subspace $\mathcal{V}$ of dimension i, let $\mathcal{S}_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V},\, \|\mathbf{y}\|_2 = 1\}$, and let

$$\mathcal{S}'_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V} \cap \mathcal{F}^{\perp},\, \|\mathbf{y}\|_2 = 1\}, \quad \text{where} \quad \mathcal{F} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}\}.$$

($\mathcal{V} \cap \mathcal{F}^{\perp} \neq \mathbf{0}$—otherwise $\dim(\mathcal{V} + \mathcal{F}^{\perp}) = \dim \mathcal{V} + \dim \mathcal{F}^{\perp} = n + 1$, which is impossible.) So $\mathcal{S}'_{\mathcal{V}}$ contains vectors of $\mathcal{S}_{\mathcal{V}}$ of the form $\mathbf{y} = (0, \ldots, 0, y_i, \ldots, y_n)^T$ with $\sum_{j=i}^{n} |y_j|^2 = 1$, and for each subspace $\mathcal{V}$ with $\dim \mathcal{V} = i$,

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=i}^{n} \lambda_j |y_j|^2 \leq \lambda_i \sum_{j=i}^{n} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}'_{\mathcal{V}}.$$

Since $\mathcal{S}'_{\mathcal{V}} \subseteq \mathcal{S}_{\mathcal{V}}$, it follows that $\min\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \min\limits_{\mathcal{S}'_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \lambda_i$, and hence

$$\max_{\mathcal{V}} \min_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \lambda_i.$$

To reverse this inequality, let $\tilde{\mathcal{V}} = span\,\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i\}$, and observe that

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=1}^{i} \lambda_j |y_j|^2 \geq \lambda_i \sum_{j=1}^{i} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}_{\tilde{\mathcal{V}}},$$

so $\max\limits_{\mathcal{V}} \min\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \max\limits_{\mathcal{S}_{\tilde{\mathcal{V}}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i$.

7.5.12. Just as before, it suffices to prove $\lambda_i = \min\limits_{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1} \in \mathcal{C}^n} \; \max\limits_{\substack{\mathbf{y} \perp \mathbf{v}_1, \ldots, \mathbf{v}_{i-1} \\ \|\mathbf{y}\|_2 = 1}} \mathbf{y}^*\mathbf{D}\mathbf{y}$. For each set $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}\}$, let $\mathcal{S}_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V}^{\perp},\, \|\mathbf{y}\|_2 = 1\}$, and let

$$\mathcal{S}'_{\mathcal{V}} = \{\mathbf{y} \in \mathcal{V}^{\perp} \cap \mathcal{T}^{\perp},\, \|\mathbf{y}\|_2 = 1\}, \quad \text{where} \quad \mathcal{T} = \{\mathbf{e}_{i+1}, \ldots, \mathbf{e}_n\}$$

($\mathcal{V}^\perp \cap \mathcal{T}^\perp \neq \mathbf{0}$—otherwise $\dim(\mathcal{V}^\perp + \mathcal{T}^\perp) = \dim \mathcal{V}^\perp + \dim \mathcal{T}^\perp = n+1$, which is impossible.) So $\mathcal{S}'_{\mathcal{V}}$ contains vectors of $\mathcal{S}_{\mathcal{V}}$ of the form $\mathbf{y} = (y_1, \ldots, y_i, 0, \ldots, 0)^T$ with $\sum_{j=1}^{i} |y_j|^2 = 1$, and for each $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}\}$,

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=1}^{i} \lambda_j |y_j|^2 \geq \lambda_i \sum_{j=1}^{i} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}'_{\mathcal{V}}.$$

Since $\mathcal{S}'_{\mathcal{V}} \subseteq \mathcal{S}_{\mathcal{V}}$, it follows that $\max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \max\limits_{\mathcal{S}'_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i$, and hence

$$\min_{\mathcal{V}} \max_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \geq \lambda_i.$$

This inequality is reversible because if $\tilde{\mathcal{V}} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{i-1}\}$, then every $\mathbf{y} \in \tilde{\mathcal{V}}$ has the form $\mathbf{y} = (0, \ldots, 0, y_i, \ldots, y_n)^T$, so

$$\mathbf{y}^*\mathbf{D}\mathbf{y} = \sum_{j=i}^{n} \lambda_j |y_j|^2 \leq \lambda_i \sum_{j=i}^{n} |y_j|^2 = \lambda_i \quad \text{for all } \mathbf{y} \in \mathcal{S}_{\tilde{\mathcal{V}}},$$

and thus $\min\limits_{\mathcal{V}} \max\limits_{\mathcal{S}_{\mathcal{V}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \max\limits_{\mathcal{S}_{\tilde{\mathcal{V}}}} \mathbf{y}^*\mathbf{D}\mathbf{y} \leq \lambda_i$. The solution for Exercise 7.5.11 can be adapted in a similar fashion to prove the alternate max-min expression.

7.5.13. (a) Unitary matrices are unitarily diagonalizable because they are normal. Furthermore, if $(\lambda, \mathbf{x})$ is an eigenpair for a unitary $\mathbf{U}$, then

$$\|\mathbf{x}\|_2^2 = \|\mathbf{U}\mathbf{x}\|_2^2 = \|\lambda\mathbf{x}\|_2^2 = |\lambda|^2 \, \|\mathbf{x}\|_2^2 \implies |\lambda| = 1 \implies \lambda = \cos\theta + \mathrm{i}\sin\theta = \mathrm{e}^{\mathrm{i}\theta}.$$

(b) This is a special case of Exercise 7.2.26 whose solution is easily adapted to provide the solution for the case at hand.

Solutions for exercises in section 7. 6

7.6.1. Check the pivots in the $\mathbf{LDL}^T$ factorization to see that $\mathbf{A}$ and $\mathbf{C}$ are positive definite. $\mathbf{B}$ is positive semidefinite.

7.6.2. (a) Examining Figure 7.6.7 shows that the force on m_1 to the left, by Hooke's law, is $F_1^{(l)} = kx_1$, and the force to the right is $F_1^{(r)} = k(x_2 - x_1)$, so the total force on m_1 is $F_1 = F_1^{(l)} - F_1^{(r)} = k(2x_1 - x_2)$. Similarly, the total force on m_2 is $F_2 = k(-x_1 + 2x_2)$. Using Newton's laws $F_1 = m_1 a_1 = m_1 x_1''$ and $F_2 = m_2 a_2 = m_2 x_2''$ yields the two second-order differential equations

$$\begin{aligned} m_1 x_1''(t) &= k(2x_1 - x_2) \\ m_2 x_2''(t) &= k(-x_1 + 2x_2) \end{aligned} \implies \mathbf{M}\mathbf{x}'' = \mathbf{K}\mathbf{x},$$

where $\mathbf{M} = \begin{pmatrix} m_1 & 0 \\ 0 & m_2 \end{pmatrix}$, and $\mathbf{K} = k \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix}$.

(b) $\lambda = (3 \pm \sqrt{3})/2$, and the normal modes are determined by the corresponding eigenvectors, which are found in the usual way by solving

$$(\mathbf{K} - \lambda \mathbf{M})\mathbf{v} = \mathbf{0}.$$

They are

$$\mathbf{v}_1 = \begin{pmatrix} -1 - \sqrt{3} \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{pmatrix} -1 + \sqrt{3} \\ 1 \end{pmatrix}$$

(c) This part is identical to that in Example 7.6.1 (p. 559) except a 2×2 matrix is used in place of a 3×3 matrix.

7.6.3. Each mass "feels" only the spring above and below it, so

$$\begin{aligned} m_1 y_1'' &= \text{Force up} - \text{Force down} = k y_1 - k(y_2 - y_1) = k(2y_1 - y_2) \\ m_2 y_2'' &= \text{Force up} - \text{Force down} = k(y_2 - y_1) - k(y_3 - y_2) = k(-y_1 + 2y_2 - y_3) \\ m_3 y_3'' &= \text{Force up} - \text{Force down} = k(y_3 - y_2) \end{aligned}$$

(b) Gerschgorin's theorem (p. 498) shows that the eigenvalues are nonnegative, as since $\det(\mathbf{K}) \neq 0$, it follows that $\mathbf{K}$ is positive definite.

(c) The same technique used in the vibrating beads problem in Example 7.6.1 (p. 559) shows that modes are determined by the eigenvectors. Some computation is required to produce $\lambda_1 \approx .198$, $\lambda_2 \approx 1.55$, and $\lambda_3 \approx 3.25$. The modes are defined by the associated eigenvectors

$$\mathbf{x}_1 = \begin{pmatrix} \gamma \\ \alpha \\ \beta \end{pmatrix} \approx \begin{pmatrix} .328 \\ .591 \\ .737 \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} -\beta \\ -\gamma \\ \alpha \end{pmatrix}, \quad \text{and} \quad \mathbf{x}_3 = \begin{pmatrix} -\alpha \\ \beta \\ -\gamma \end{pmatrix}.$$

7.6.4. Write the quadratic form as $13x^2 + 10xy + 13y^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 13 & 5 \\ 5 & 13 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{z}^T \mathbf{A} \mathbf{z}$. We know from Example 7.6.3 on p. 567 that if $\mathbf{Q}$ is an orthogonal matrix such that $\mathbf{Q}^T \mathbf{A} \mathbf{Q} = \mathbf{D} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, and if $\mathbf{w} = \mathbf{Q}^T \mathbf{z} = \begin{pmatrix} u \\ v \end{pmatrix}$, then

$$13x^2 + 10xy + 13y^2 = \mathbf{z}^T \mathbf{A} \mathbf{z} = \mathbf{w}^T \mathbf{D} \mathbf{w} = \lambda_1 u^2 + \lambda_2 v^2.$$

Computation reveals that $\lambda_1 = 8$, $\lambda_2 = 18$, and $\mathbf{Q} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$, so the graph of $13x^2 + 10xy + 13y^2 = 72$ is the same as that for $18u^2 + 8v^2 = 72$ or, equivalently, $u^2/9 + v^2/4 = 1$. It follows from (5.6.13) on p. 326 that the uv-coordinate system results from rotating the standard xy-coordinate system counterclockwise by $45°$.

7.6.5. Since $\mathbf{A}$ is symmetric, the LDU factorization is really $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$ (see Exercise 3.10.9 on p. 157). In other words, $\mathbf{A} \cong \mathbf{D}$, so Sylvester's law of inertia guarantees that the inertia of $\mathbf{A}$ is the same as the inertia of $\mathbf{D}$.

7.6.6. (a) Notice that, in general, when $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is expanded, the coefficient of $x_i x_j$ is given by $(a_{ij} + a_{ji})/2$. Therefore, for the problem at hand, we can take

$$\mathbf{A} = \frac{1}{9}\begin{pmatrix} -2 & 2 & 8 \\ 2 & 7 & 10 \\ 8 & 10 & 4 \end{pmatrix}.$$

(b) Gaussian elimination provides $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$, where

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & 2 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \begin{pmatrix} -2/9 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

so the inertia of $\mathbf{A}$ is $(1, 1, 1)$. Setting $\mathbf{y} = \mathbf{L}^T\mathbf{x}$ (or, $\mathbf{x} = (\mathbf{L}^T)^{-1}\mathbf{y}$) yields

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{D}\mathbf{y} = -\frac{2}{9}y_1^2 + y_2^2.$$

(c) No, the form is indefinite.

(d) The eigenvalues of $\mathbf{A}$ are $\{2, -1, 0\}$, and hence the inertia is $(1, 1, 1)$.

7.6.7. $\mathbf{A}\mathbf{A}^*$ is positive definite (because $\mathbf{A}$ is nonsingular), so its eigenvalues λ_i are real and positive. Consequently, the spectral decomposition (p. 517) allows us to write $\mathbf{A}\mathbf{A}^* = \sum_{i=1}^k \lambda_i \mathbf{G}_i$. Use the results on (p. 526), and set

$$\mathbf{R} = (\mathbf{A}\mathbf{A}^*)^{1/2} = \sum_{i=1}^k \lambda_i^{1/2}\mathbf{G}_i, \quad \text{and} \quad \mathbf{R}^{-1} = (\mathbf{A}\mathbf{A}^*)^{-1/2} = \sum_{i=1}^k \lambda_i^{-1/2}\mathbf{G}_i.$$

It now follows that $\mathbf{R}$ is positive definite, and $\mathbf{A} = \mathbf{R}(\mathbf{R}^{-1}\mathbf{A}) = \mathbf{R}\mathbf{U}$, where $\mathbf{U} = \mathbf{R}^{-1}\mathbf{A}$. Finally, $\mathbf{U}$ is unitary because

$$\mathbf{U}\mathbf{U}^* = (\mathbf{A}\mathbf{A}^*)^{-1/2}\mathbf{A}\mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1/2} = \mathbf{I}.$$

If $\mathbf{R}_1\mathbf{U}_1 = \mathbf{A} = \mathbf{R}_2\mathbf{U}_2$, then $\mathbf{R}_2^{-1}\mathbf{R}_1 = \mathbf{U}_2\mathbf{U}_1^*$, which is unitary, so

$$\mathbf{R}_2^{-1}\mathbf{R}_1\mathbf{R}_1\mathbf{R}_2^{-1} = \mathbf{I} \implies \mathbf{R}_1^2 = \mathbf{R}_2^2 \implies \mathbf{R}_1 = \mathbf{R}_2 \text{ (because the } \mathbf{R}_i\text{'s are pd).}$$

7.6.8. The 2-norm condition number is the ratio of the largest to smallest singular values. Since $\mathbf{L}$ is symmetric and positive definite, the singular values are the eigenvalues, and, by (7.6.8), $\max \lambda_{ij} \to 8$ and $\min \lambda_{ij} \to 0$ as $n \to \infty$.

7.6.9. The procedure is essentially identical to that in Example 7.6.2. The only difference is that when (7.6.6) is applied, the result is

$$\frac{-4u_{ij} + (u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1})}{h^2} + O(h^2) = f_{ij}$$

or, equivalently,

$$4u_{ij}-(u_{i-1,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1})+O(h^4) = -h^2 f_{ij} \quad \text{for} \quad i,j = 1,2,\ldots,n.$$

If the $O(h^4)$ terms are neglected, and if the boundary values g_{ij} are taken to the right-hand side, then, with the same ordering as indicated in Example 7.6.2, the system $\mathbf{Lu} = \mathbf{g} - h^2\mathbf{f}$ is produced.

7.6.10. $$\mathbf{I}_n \otimes \mathbf{A}_n = \begin{pmatrix} \mathbf{A}_n & 0 & \cdots & 0 \\ 0 & \mathbf{A}_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_n \end{pmatrix}, \quad \mathbf{A}_n \otimes \mathbf{I}_n = \begin{pmatrix} 2\mathbf{I}_n & -\mathbf{I}_n & & & \\ -\mathbf{I}_n & 2\mathbf{I}_n & -\mathbf{I}_n & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I}_n & 2\mathbf{I}_n & -\mathbf{I}_n \\ & & & -\mathbf{I}_n & 2\mathbf{I}_n \end{pmatrix},$$

$$\mathbf{A}_n + 2\mathbf{I}_n = \mathbf{T}_n = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix}_{n\times n}, \text{ so}$$

$$(\mathbf{I}_n \otimes \mathbf{A}_n) + (\mathbf{A}_n \otimes \mathbf{I}_n) = \begin{pmatrix} \mathbf{T}_n & -\mathbf{I}_n & & & \\ -\mathbf{I}_n & \mathbf{T}_n & -\mathbf{I}_n & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathbf{I}_n & \mathbf{T}_n & -\mathbf{I}_n \\ & & & -\mathbf{I}_n & \mathbf{T}_n \end{pmatrix} = \mathbf{L}_{n^2\times n^2}.$$

Solutions for exercises in section 7. 7

7.7.1. No. This can be deduced directly from the definition of index given on p. 395, or it can be seen by looking at the Jordan form (7.7.6) on p. 579.

7.7.2. Since the index k of a 4×4 nilpotent matrix cannot exceed 4, consider the different possibilities for $k = 1,2,3,4$. For $k = 1$, $N = \mathbf{0}_{4\times 4}$ is the only possibility. If $k = 2$, the largest Jordan block in $\mathbf{N}$ is 2×2, so

$$\mathbf{N} = \left(\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right) \quad \text{and} \quad \mathbf{N} = \left(\begin{array}{cc|c|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right)$$

are the only possibilities. If $k = 3$ or $k = 4$, then the largest Jordan block is 3×3 or 4×4, so

$$\mathbf{N} = \left(\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right) \quad \text{and} \quad \mathbf{N} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

are the only respective possibilities.

7.7.3. Let $k = \mathit{index}(\mathbf{L})$, and let ζ_i denote the number of blocks of size $i \times i$ or larger. This number is determined by the number of chains of length i or larger, and such chains emanate from the vectors in $\mathcal{S}_{k-1} \cup \mathcal{S}_{k-2} \cup \cdots \cup \mathcal{S}_{i-1} = \mathcal{B}_{i-1}$. Since $\mathcal{B}_{i-1}$ is a basis for $\mathcal{M}_{i-1}$, it follows that $\zeta_i = \dim \mathcal{M}_{i-1} = r_{i-1} - r_i$, where $r_i = \mathit{rank}\left(\mathbf{L}^i\right)$ —recall (7.7.3).

7.7.4. $\mathbf{x} \in \mathcal{M}_i = R\left(\mathbf{L}^i\right) \cap N(\mathbf{L}) \implies \mathbf{x} = \mathbf{L}^i\mathbf{y}$ for some $\mathbf{y}$ and $\mathbf{L}\mathbf{x} = \mathbf{0} \implies \mathbf{x} = \mathbf{L}^{i-1}(\mathbf{L}\mathbf{y})\mathbf{y}$ and $\mathbf{L}\mathbf{x} = \mathbf{0} \implies \mathbf{x} \in R\left(\mathbf{L}^{i-1}\right) \cap N(\mathbf{L}) = \mathcal{M}_{i-1}$.

7.7.5. It suffices to prove that $R\left(\mathbf{L}^{k-1}\right) \subseteq N(\mathbf{L})$, and this is accomplished by writing

$$\mathbf{x} \in R\left(\mathbf{L}^{k-1}\right) \Longrightarrow \mathbf{x} = \mathbf{L}^{k-1}\mathbf{y} \text{ for some } \mathbf{y} \Longrightarrow \mathbf{L}\mathbf{x} = \mathbf{L}^k\mathbf{y} = \mathbf{0} \Longrightarrow \mathbf{x} \in N(\mathbf{L}).$$

7.7.6. This follows from the result on p. 211.

7.7.7. $\mathbf{L}^2 = \mathbf{0}$ means that $\mathbf{L}$ is nilpotent of index $k = 2$. Consequently, the size of the largest Jordan block in $\mathbf{N}$ is 2×2. Since $r_1 = 2$ and $r_i = 0$ for $i \geq 2$, the number of 2×2 blocks is $r_1 - 2r_2 + r_3 = 2$, so the Jordan form is

$$\mathbf{N} = \left(\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

In this case, $\mathcal{M}_0 = N(\mathbf{L}) = R(\mathbf{L}) = \mathcal{M}_1$ because $\mathbf{L}^2 = \mathbf{0} \implies R(\mathbf{L}) \subseteq N(\mathbf{L})$ and $\dim R(\mathbf{L}) = 2 = \dim R(\mathbf{L})$. Now, $\mathcal{S}_1 = \{\mathbf{L}_{*1}, \mathbf{L}_{*2}\}$ (the basic columns in $\mathbf{L}$) is a basis for $\mathcal{M}_1 = R(\mathbf{L})$, and $\mathcal{S}_0 = \phi$. Since $\mathbf{x}_1 = \mathbf{e}_1$ and $\mathbf{x}_2 = \mathbf{e}_2$ are solutions for $\mathbf{L}\mathbf{x}_1 = \mathbf{L}_{*1}$ and $\mathbf{L}\mathbf{x}_2 = \mathbf{L}_{*1}$, respectively, there are two Jordan chains, namely $\{\mathbf{L}\mathbf{x}_1, \mathbf{x}_1\} = \{\mathbf{L}_{*1}, \mathbf{e}_1\}$ and $\{\mathbf{L}\mathbf{x}_2, \mathbf{x}_2\} = \{\mathbf{L}_{*2}, \mathbf{e}_2\}$, so

$$\mathbf{P} = [\,\mathbf{L}_{*1} \,|\, \mathbf{e}_1 \,|\, \mathbf{L}_{*2} \,|\, \mathbf{e}_2\,] = \begin{pmatrix} 3 & 1 & 3 & 0 \\ -2 & 0 & -1 & 1 \\ 1 & 0 & -1 & 0 \\ -5 & 0 & -4 & 0 \end{pmatrix}.$$

Use direct computation to verify that $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$.

7.7.8. Computing $r_i = \mathit{rank}\left(\mathbf{L}^i\right)$ reveals that $r_1 = 4$, $r_2 = 1$, and $r_3 = 0$, so the index of $\mathbf{L}$ is $k = 3$, and

$$\begin{aligned} \text{the number of } 3 \times 3 \text{ blocks } &= r_2 - 2r_3 + r_4 = 1, \\ \text{the number of } 2 \times 2 \text{ blocks } &= r_1 - 2r_2 + r_3 = 2, \\ \text{the number of } 1 \times 1 \text{ blocks } &= r_0 - 2r_1 + r_2 = 1. \end{aligned}$$

Consequently, the Jordan form of $\mathbf{L}$ is

$$\mathbf{N} = \left(\begin{array}{ccc|cc|cc|c} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Notice that four Jordan blocks were found, and this agrees with the fact that $\dim N(\mathbf{L}) = 8 - rank(\mathbf{L}) = 4$.

7.7.9. If $\mathbf{N}_i$ is an $n_i \times n_i$, nilpotent block as described in (7.7.5), and if $\mathbf{D}_i$ is the diagonal matrix $\mathbf{D}_i = \operatorname{diag}\left(1, \epsilon_i, \ldots, \epsilon^{n_i-1}\right)$, then $\mathbf{D}_i^{-1}\mathbf{N}_i\mathbf{D}_i = \epsilon_i\mathbf{N}_i$. Therefore, if $\mathbf{P}^{-1}\mathbf{L}\mathbf{P} = \mathbf{N}$ is in Jordan form, and if $\mathbf{Q} = \mathbf{P}\mathbf{D}$, where $\mathbf{D}$ is the block-diagonal matrix $\mathbf{D} = \operatorname{diag}(\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_t)$, then $\mathbf{Q}^{-1}\mathbf{L}\mathbf{Q} = \tilde{\mathbf{N}}$.

Solutions for exercises in section 7. 8

7.8.1. Since $rank(\mathbf{A}) = 7$, $rank\left(\mathbf{A}^2\right) = 6$, and $rank\left(\mathbf{A}^3\right) = 5 = rank\left(\mathbf{A}^{3+i}\right)$, there is one 3×3 Jordan block associates with $\lambda = 0$. Since $rank(\mathbf{A} + \mathbf{I}) = 6$ and $rank\left((\mathbf{A} + \mathbf{I})^2\right) = 5 = rank\left((\mathbf{A} + \mathbf{I})^{2+i}\right)$, there is one 1×1 and one 2×2 Jordan block associated with $\lambda = -1$. Finally, $rank(\mathbf{A} - \mathbf{I}) = rank\left((\mathbf{A} - \mathbf{I})^{1+i}\right)$ implies there are two 1×1 blocks associated with $\lambda = 1$—i.e., $\lambda = 1$ is a semisimple eigenvalue. Therefore, the Jordan form for $\mathbf{A}$ is

$$\mathbf{J} = \left(\begin{array}{ccc|cc|c|c|c} 0 & 1 & 0 & & & & & \\ & 0 & 1 & & & & & \\ & & 0 & & & & & \\ \hline & & & -1 & 1 & & & \\ & & & 0 & -1 & & & \\ \hline & & & & & -1 & & \\ \hline & & & & & & 1 & \\ \hline & & & & & & & 1 \end{array}\right).$$

7.8.2. As noted in Example 7.8.3, $\sigma(\mathbf{A}) = \{1\}$ and $k = index(1) = 2$. Use the procedure on p. 211 to determine a basis for $\mathcal{M}_{k-1} = \mathcal{M}_1 = R(\mathbf{A} - \mathbf{I}) \cap N(\mathbf{A} - \mathbf{I})$ to be $\mathcal{S}_1 = \left\{\begin{pmatrix} 1 \\ -2 \\ -2 \end{pmatrix} = \mathbf{b}_1\right\}$. (You might also determine this just by inspection.) A basis for $N(\mathbf{A} - \mathbf{I})$ is easily found to be $\left\{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}\right\}$, so examining

the basic columns of $\begin{pmatrix} 1 & 0 & 1 \\ -2 & 1 & 0 \\ -2 & 0 & -2 \end{pmatrix}$ yields the extension set $\mathcal{S}_0 = \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{b}_2 \right\}$. Solving $(\mathbf{A} - \mathbf{I})\mathbf{x} = \mathbf{b}_1$ produces $\mathbf{x} = \mathbf{e}_3$, so the associated Jordan chain is $\{\mathbf{b}_1, \mathbf{e}_3\}$, and thus $\mathbf{P} = \left(\mathbf{b}_1 \,|\, \mathbf{e}_3 \,|\, \mathbf{b}_2\right) = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 0 & 1 \\ -2 & 1 & 0 \end{pmatrix}$. It's easy to check that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ is indeed the Jordan form for $\mathbf{A}$.

7.8.3. If $k = \mathit{index}(\lambda)$, then the size of the largest Jordan block associated with λ is $k \times k$. This insures that λ must be repeated at least k times, and thus $\mathit{index}(\lambda) \leq \mathit{alg\ mult}(\lambda)$.

7.8.4. $\mathit{index}(\lambda) = 1$ if and only if every Jordan block is 1×1, which happens if and only if the number of eigenvectors associated with λ in $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$ is the same as the number of Jordan blocks, and this is just another way to say that $\mathit{alg\ mult}_{\mathbf{A}}(\lambda) = \mathit{geo\ mult}_{\mathbf{A}}(\lambda)$, which is the definition of λ being a semisimple eigenvalue (p. 510).

7.8.5. Notice that $\mathbf{R}^2 = \mathbf{I}$, so $\mathbf{R}^{-1} = \mathbf{R} = \mathbf{R}^T$, and if $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix}$ is a generic Jordan block, then $\mathbf{R}^{-1}\mathbf{J}_\star\mathbf{R} = \mathbf{R}\mathbf{J}_\star\mathbf{R} = \mathbf{J}_\star^T$. Thus every Jordan block is similar to its transpose. Given any Jordan form, reversal matrices of appropriate sizes can be incorporated into a block-diagonal matrix $\widetilde{\mathbf{R}}$ such that $\widetilde{\mathbf{R}}^{-1}\mathbf{J}\widetilde{\mathbf{R}} = \mathbf{J}^T$ showing that $\mathbf{J}$ is similar to $\mathbf{J}^T$. Consequently, if $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, then

$$\mathbf{A}^T = \mathbf{P}^{-1^T}\mathbf{J}^T\mathbf{P}^T = \mathbf{P}^{-1^T}\widetilde{\mathbf{R}}^{-1}\mathbf{J}\widetilde{\mathbf{R}}\mathbf{P}^T = \mathbf{P}^{-1^T}\widetilde{\mathbf{R}}^{-1}\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\widetilde{\mathbf{R}}\mathbf{P}^T = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q},$$

where $\mathbf{Q} = \mathbf{P}\widetilde{\mathbf{R}}\mathbf{P}^T$.

7.8.6. If $\sigma(\mathbf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_s\}$, where $\mathit{alg\ mult}(\lambda_i) = a_i$, then the characteristic equation for $\mathbf{A}$ is $0 = (x - \lambda_1)^{a_1}(x - \lambda_2)^{a_2} \cdots (x - \lambda_s)^{a_s} = c(x)$. If

$$\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \text{ is in Jordan form with } \mathbf{J}_\star = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda_i \end{pmatrix}$$

representing a generic Jordan block, then

$$c(\mathbf{A}) = c(\mathbf{P}\mathbf{J}\mathbf{P}^{-1}) = \mathbf{P}c(\mathbf{J})\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \ddots & & \\ & c(\mathbf{J}_\star) & \\ & & \ddots \end{pmatrix}\mathbf{P}^{-1}.$$

Notice that if $\mathbf{J}_\star$ is $r \times r$, then $(\mathbf{J}_\star - \lambda_i\mathbf{I})^r = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & & 0 \end{pmatrix}^r = \mathbf{0}$. Since the size of the largest Jordan block associated with λ_i is $k_i \times k_i$, where $k_i = \mathit{index}(\lambda_i) \leq \mathit{alg\ mult}(\lambda_i) = a_i$, it follows that $(\mathbf{J}_\star - \lambda_i\mathbf{I})^{a_i} = \mathbf{0}$. Consequently $c(\mathbf{J}_\star) = \mathbf{0}$ for every Jordan block, and thus $c(\mathbf{A}) = \mathbf{0}$.

7.8.7. By using the Jordan form for $\mathbf{A}$, one can find a similarity transformation $\mathbf{P}$ such that $\mathbf{P}^{-1}(\mathbf{A}-\lambda\mathbf{I})\mathbf{P} = \begin{pmatrix} \mathbf{L}_{m\times m} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$ with $\mathbf{L}^k = \mathbf{0}$ and $\mathbf{C}$ nonsingular. Therefore, $\mathbf{P}^{-1}(\mathbf{A}-\lambda\mathbf{I})^k\mathbf{P} = \begin{pmatrix} \mathbf{0}_{m\times m} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^k \end{pmatrix}$, and thus

$$\dim N\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = n - rank\big((\mathbf{A}-\lambda\mathbf{I})^k\big) = n - rank\big(\mathbf{C}^k\big) = m.$$

It is also true that $\dim N\big((\mathbf{A}-\lambda\mathbf{I})^m\big) = m$ because the nullspace remains the same for all powers beyond the index (p. 395).

7.8.8. To prove $\mathcal{M}_{k_j}(\lambda_j) = \mathbf{0}$, suppose $\mathbf{x} \in \mathcal{M}_{k_j}(\lambda_j)$ so that $(\mathbf{A}-\lambda_j\mathbf{I})\mathbf{x} = \mathbf{0}$ and $\mathbf{x} = (\mathbf{A}-\lambda_j\mathbf{I})^{k_j}\mathbf{z}$ for some $\mathbf{z}$. Combine these with the properties of $index(\lambda_j)$ (p. 587) to obtain

$$(\mathbf{A}-\lambda_j\mathbf{I})^{k_j+1}\mathbf{z} = \mathbf{0} \implies (\mathbf{A}-\lambda_j\mathbf{I})^{k_j}\mathbf{z} = \mathbf{0} \implies \mathbf{x} = \mathbf{0}.$$

The fact that the subspaces are nested follows from the observation that if $\mathbf{x} \in \mathcal{M}_{i+1}(\lambda_j)$, then $\mathbf{x} = (\mathbf{A}-\lambda_j\mathbf{I})^{i+1}\mathbf{z}$ and $(\mathbf{A}-\lambda_j\mathbf{I})\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = (\mathbf{A}-\lambda_j\mathbf{I})^i\big((\mathbf{A}-\lambda_j\mathbf{I})\mathbf{z}\big)$ and $(\mathbf{A}-\lambda_j\mathbf{I})\mathbf{x} = \mathbf{0}$, so $\mathcal{M}_{i+1}(\lambda_j) \subseteq \mathcal{M}_i(\lambda_j)$.

7.8.9. $\mathbf{b}(\lambda_j) \in \mathcal{S}_i(\lambda_j) \subseteq \mathcal{M}_i(\lambda_j) = R\big((\mathbf{A}-\lambda_j\mathbf{I})^i\big) \cap N(\mathbf{A}-\lambda_j\mathbf{I}) \subseteq R\big((\mathbf{A}-\lambda_j\mathbf{I})^i\big)$.

7.8.10. No—consider $\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$ and $\lambda = 1$.

7.8.11. (a) All of these facts are established by straightforward arguments using elementary properties of matrix algebra, so the details are omitted here.

(b) To show that the eigenvalues of $\mathbf{A}\otimes\mathbf{B}$ are $\{\lambda_i\mu_j\}_{i=1}^{m}{}_{j=1}^{n}$, let $\mathbf{J}_A = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ and $\mathbf{J}_B = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}$ be the respective Jordan forms for $\mathbf{A}$ and $\mathbf{B}$, and use properties from (a) to establish that $\mathbf{A}\otimes\mathbf{B}$ is similar to $\mathbf{J}_A\otimes\mathbf{J}_B$ by writing

$$\begin{aligned}\mathbf{J}_A\otimes\mathbf{J}_B &= (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})\otimes(\mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}) = (\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1})(\mathbf{A}\otimes\mathbf{B})(\mathbf{P}\otimes\mathbf{Q}) \\ &= (\mathbf{P}\otimes\mathbf{Q})^{-1}(\mathbf{A}\otimes\mathbf{B})(\mathbf{P}\otimes\mathbf{Q})\end{aligned}$$

Thus the eigenvalues of $\mathbf{A}\otimes\mathbf{B}$ are the same as those of $\mathbf{J}_A\otimes\mathbf{J}_B$, and because $\mathbf{J}_A$ and $\mathbf{J}_B$ are upper triangular with the λ_i's and μ_i's on the diagonal, it's clear that $\mathbf{J}_A\otimes\mathbf{J}_B$ is also upper triangular with diagonal entries being $\lambda_i\mu_j$.

To show that the eigenvalues of $(\mathbf{A}\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{B})$ are $\{\lambda_i+\mu_j\}_{i=1}^{m}{}_{j=1}^{n}$, show that $(\mathbf{A}\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{B})$ is similar to $(\mathbf{J}_A\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{J}_B)$ by writing

$$\begin{aligned}(\mathbf{J}_A\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{J}_B) &= (\mathbf{P}^{-1}\mathbf{A}\mathbf{P})\otimes(\mathbf{Q}^{-1}\mathbf{I}\mathbf{Q}) + (\mathbf{P}^{-1}\mathbf{I}\mathbf{P})\otimes(\mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}) \\ &= (\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1})(\mathbf{A}\otimes\mathbf{I})(\mathbf{P}\otimes\mathbf{Q}) \\ &\qquad + (\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1})(\mathbf{I}\otimes\mathbf{B})(\mathbf{P}\otimes\mathbf{Q}) \\ &= (\mathbf{P}^{-1}\otimes\mathbf{Q}^{-1})\big[(\mathbf{A}\otimes\mathbf{I})+(\mathbf{I}\otimes\mathbf{B})\big](\mathbf{P}\otimes\mathbf{Q}) \\ &= (\mathbf{P}\otimes\mathbf{Q})^{-1}\big[(\mathbf{A}\otimes\mathbf{I})+(\mathbf{I}\otimes\mathbf{B})\big](\mathbf{P}\otimes\mathbf{Q}).\end{aligned}$$

Thus $(\mathbf{A}\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{B})$ and $(\mathbf{J}_A\otimes\mathbf{I}_n)+(\mathbf{I}_m\otimes\mathbf{J}_B)$ have the same eigenvalues, and the latter matrix is easily seen to be an upper-triangular matrix whose diagonal entries are $\{\lambda_i+\mu_j\}_{i=1}^{m}{}_{j=1}^{n}$.

7.8.12. It was established in Exercise 7.6.10 (p. 573) that $\mathbf{L}_{n^2\times n^2}=(\mathbf{I}_n\otimes\mathbf{A}_n)+(\mathbf{A}_n\otimes\mathbf{I}_n)$, where

$$\mathbf{A}_n=\begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{n\times n}$$

is the finite difference matrix of Example 1.4.1 (p. 19). The eigenvalues of $\mathbf{A}_n$ were determined in Exercise 7.2.18 (p. 522) to be $\mu_j=4\sin^2[j\pi/2(n+1)]$ for $j=1,2,\ldots,n$, so it follows from the last property in Exercise 7.8.11 that the n^2 eigenvalues of $\mathbf{L}_{n^2\times n^2}=(\mathbf{I}_n\otimes\mathbf{A}_n)+(\mathbf{A}_n\otimes\mathbf{I}_n)$ are

$$\lambda_{ij}=\mu_i+\mu_j=4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right)+\sin^2\left(\frac{j\pi}{2(n+1)}\right)\right],\qquad i,j=1,2,\ldots,n.$$

7.8.13. The same argument given in the solution of the last part of Exercise 7.8.11 applies to show that if $\mathbf{J}$ is the Jordan form for $\mathbf{A}$, then $\mathbf{L}$ is similar to $(\mathbf{I}\otimes\mathbf{I}\otimes\mathbf{J})+(\mathbf{I}\otimes\mathbf{J}\otimes\mathbf{I})+(\mathbf{J}\otimes\mathbf{I}\otimes\mathbf{I})$, and since $\mathbf{J}$ is upper triangular with the eigenvalues $\mu_j=4\sin^2[j\pi/2(n+1)]$ of $\mathbf{A}$ (recall Exercise 7.2.18 (p. 522)) on the diagonal of $\mathbf{J}$, it follows that the eigenvalues of $\mathbf{L}_{n^3\times n^3}$ are the n^3 numbers

$$\lambda_{ijk}=\mu_i+\mu_j+\mu_k=4\left[\sin^2\left(\frac{i\pi}{2(n+1)}\right)+\sin^2\left(\frac{j\pi}{2(n+1)}\right)+\sin^2\left(\frac{k\pi}{2(n+1)}\right)\right]$$

for $i,j,k=1,2,\ldots,n$.

Solutions for exercises in section 7. 9

7.9.1. If $u_i(t)$ denotes the number of pounds of pollutant in lake i at time $t>0$, then the concentration of pollutant in lake i at time t is $u_i(t)/V$ lbs/gal, so the model $u_i'(t)=$ (lbs/sec) coming in$-$(lbs/sec) going out produces the system

$$\begin{aligned} u_1' &= \frac{4r}{V}u_2-\frac{4r}{V}u_1 \\ u_2' &= \frac{2r}{V}u_1+\frac{3r}{V}u_3-\frac{5r}{V}u_2 \\ u_3' &= \frac{2r}{V}u_1+\frac{r}{V}u_2-\frac{3r}{V}u_3 \end{aligned}\quad\text{or}\quad \begin{pmatrix} u_1' \\ u_2' \\ u_3' \end{pmatrix}=\frac{r}{V}\begin{pmatrix} -4 & 4 & 0 \\ 2 & -5 & 3 \\ 2 & 1 & -3 \end{pmatrix}\begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix}.$$

The solution of $\mathbf{u}'=\mathbf{A}\mathbf{u}$ with $\mathbf{u}(0)=\mathbf{c}$ is $\mathbf{u}=\mathrm{e}^{\mathbf{A}t}\mathbf{c}$. The eigenvalues of $\mathbf{A}$ are $\lambda_1=0$ with *alg mult*$(\lambda_1)=1$ and $\lambda_2=-6r/V$ with *index*$(\lambda_2)=2$, so

$$\mathbf{u}=\mathrm{e}^{\mathbf{A}t}\mathbf{c}=\mathrm{e}^{\lambda_1 t}\mathbf{G}_1\mathbf{c}+\mathrm{e}^{\lambda_2 t}\mathbf{G}_2\mathbf{c}+t\mathrm{e}^{\lambda_2 t}(\mathbf{A}-\lambda_2\mathbf{I})\mathbf{G}_2\mathbf{c}.$$

Since $\lambda_1 = 0$ is a simple eigenvalue, it follows from (7.2.12) on p. 518 that $\mathbf{G}_1 = \mathbf{x}\mathbf{y}^T/\mathbf{y}^T\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{y}^T$ are any pair of respective right-hand and left-hand eigenvectors associated with $\lambda_1 = 0$. By observing that $\mathbf{A}\mathbf{e} = \mathbf{0}$ and $\mathbf{e}^T\mathbf{A} = \mathbf{0}^T$ for $\mathbf{e} = (1,\, 1,\, 1)^T$ (this is a result of being a closed system), and by using $\mathbf{G}_1 + \mathbf{G}_2 = \mathbf{I}$, we have (by using $\mathbf{x} = \mathbf{y} = \mathbf{e}$)

$$\mathbf{G}_1 = \frac{\mathbf{e}\mathbf{e}^T}{3}, \quad \mathbf{G}_2 = \mathbf{I} - \frac{\mathbf{e}\mathbf{e}^T}{3}, \quad \text{and} \quad (\mathbf{A} - \lambda_2\mathbf{I})\mathbf{G}_2 = \mathbf{A} - \lambda_2\mathbf{I} + \frac{\lambda_2}{3}\mathbf{e}\mathbf{e}^T.$$

If $\alpha = (c_1 + c_2 + c_3)/3 = \mathbf{e}^T\mathbf{c}/3$ denotes the average of the initial values, then $\mathbf{G}_1\mathbf{c} = \alpha\mathbf{e}$ and $\mathbf{G}_2\mathbf{c} = \mathbf{c} - \alpha\mathbf{e}$, so

$$\mathbf{u}(t) = \alpha\mathbf{e} + \mathrm{e}^{\lambda_2 t}(\mathbf{c} - \alpha\mathbf{e}) + t\mathrm{e}^{\lambda_2 t}\big[\mathbf{A}\mathbf{c} - \lambda_2(\mathbf{c} - \alpha\mathbf{e})\big] \quad \text{for} \quad \lambda_2 = -6r/V.$$

Since $\lambda_2 < 0$, it follows that $\mathbf{u}(t) \to \alpha\mathbf{e}$ as $t \to \infty$. In other words, the long-run amount of pollution in each lake is the same—namely α lbs—and this is what common sense would dictate.

7.9.2. It follows from (7.9.9) that $f_i(\mathbf{A}) = \mathbf{G}_i$.

7.9.3. We know from Exercise 7.9.2 that $\mathbf{G}_i = f_i(\mathbf{A})$ for $f_i(z) = \begin{cases} 1 & \text{when } z = \lambda_i, \\ 0 & \text{otherwise,} \end{cases}$ and from Example 7.9.4 (p. 606) we know that every function of $\mathbf{A}$ is a polynomial in $\mathbf{A}$.

7.9.4. Using $f(z) = z^k$ in (7.9.9) on p. 603 produces the desired result.

7.9.5. Using $f(z) = z^n$ in (7.9.2) on p. 600 produces the desired result.

7.9.6. $\mathbf{A}$ is the matrix in Example 7.9.2, so the results derived there imply that

$$\mathrm{e}^{\mathbf{A}} = \mathrm{e}^2\mathbf{G}_1 + \mathrm{e}^4\mathbf{G}_2 + \mathrm{e}^4(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2 = \begin{pmatrix} 3\mathrm{e}^4 & 2\mathrm{e}^4 & \mathrm{e}^4 - \mathrm{e}^2 \\ -2\mathrm{e}^4 & -\mathrm{e}^4 & -4\mathrm{e}^4 - 2\mathrm{e}^2 \\ 0 & 0 & \mathrm{e}^2 \end{pmatrix}.$$

7.9.7. The eigenvalues of $\mathbf{A}$ are $\lambda_1 = 1$ and $\lambda_2 = 4$ with *alg mult*$(1) = 1$ and *index*$(4) = 2$, so

$$f(\mathbf{A}) = f(1)\mathbf{G}_1 + f(4)\mathbf{G}_2 + f'(4)(\mathbf{A} - 4\mathbf{I})\mathbf{G}_2$$

Since $\lambda_1 = 1$ is a simple eigenvalue, it follows from formula (7.2.12) on p. 518 that $\mathbf{G}_1 = \mathbf{x}\mathbf{y}^T/\mathbf{y}^T\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{y}^T$ are any pair of respective right-hand and left-hand eigenvectors associated with $\lambda_1 = 1$. Using $\mathbf{x} = (-2,\, 1,\, 0)^T$ and $\mathbf{y} = (1,\, 1,\, 1)^T$ produces

$$\mathbf{G}_1 = \begin{pmatrix} 2 & 2 & 2 \\ -1 & -1 & -1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{G}_2 = \mathbf{I} - \mathbf{G}_1 = \begin{pmatrix} -1 & -2 & -2 \\ 1 & 2 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Therefore,

$$f(\mathbf{A}) = 4\sqrt{\mathbf{A}} - \mathbf{I} = 3\mathbf{G}_1 + 7\mathbf{G}_2 + (\mathbf{A} - 4\mathbf{I})\mathbf{G}_2 = \begin{pmatrix} -2 & -10 & -11 \\ 6 & 15 & 10 \\ -1 & -2 & 4 \end{pmatrix}.$$

7.9.8. (a) The only point at which derivatives of $f(z) = z^{1/2}$ fail to exist are at $\lambda = 0$, so as long as $\mathbf{A}$ is nonsingular, $f(\mathbf{A}) = \sqrt{\mathbf{A}}$ is defined.

(b) If $\mathbf{A}$ is singular so that $0 \in \sigma(\mathbf{A})$ it's clear from (7.9.9) that $\mathbf{A}^{1/2}$ exists if and only if derivatives of $f(z) = z^{1/2}$ need not be evaluated at $\lambda = 0$, and this is the case if and only if $index(0) = 1$.

7.9.9. If $\mathbf{0} \neq \mathbf{x}_h \in N(\mathbf{A} - \lambda_h \mathbf{I})$, then (7.9.11) guarantees that

$$\mathbf{G}_i \mathbf{x}_h = \begin{cases} \mathbf{0} & \text{if } h \neq i, \\ \mathbf{x}_h & \text{if } h = i, \end{cases}$$

so (7.9.9) can be used to conclude that $f(\mathbf{A})\mathbf{x}_h = f(\lambda_h)\mathbf{x}_h$. It's an immediate consequence of (7.9.3) that $alg\ mult_{\mathbf{A}}(\lambda) = alg\ mult_{f(\mathbf{A})}(f(\lambda))$.

7.9.10. (a) If $\mathbf{A}_{k\times k}$ (with $k > 1$) is a Jordan block associated with $\lambda = 0$, and if $f(z) = z^k$, then $f(\mathbf{A}) = \mathbf{0}$ is not similar to $\mathbf{A} \neq \mathbf{0}$.

(b) Also, $geo\ mult_{\mathbf{A}}(0) = 1$ but $geo\ mult_{f(\mathbf{A})}(f(0)) = geo\ mult_{\mathbf{0}}(0) = k$.

(c) And $index_{\mathbf{A}}(0) = k$ while $index_{f(\mathbf{A})}(f(\lambda)) = index_{\mathbf{0}}(0) = 1$.

7.9.11. This follows because, as explained in Example 7.9.4 (p. 606), there is always a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$, and $\mathbf{A}$ commutes with $p(\mathbf{A})$.

7.9.12. Because every square matrix is similar to its transpose (recall Exercise 7.8.5 on p. 596), and because similar matrices have the same Jordan structure, transposition doesn't change the eigenvalues or their indicies. So $f(\mathbf{A})$ exists if and only if $f(\mathbf{A}^T)$ exists. As proven in Example 7.9.4 (p. 606), there is a polynomial $p(z)$ such that $f(\mathbf{A}) = p(\mathbf{A})$, so $\big[f(\mathbf{A})\big]^T = \big[p(\mathbf{A})\big]^T = p(\mathbf{A}^T) = f(\mathbf{A}^T)$. While transposition doesn't change eigenvalues, conjugate transposition does—it conjugates them—so it's possible that f can exist at $\mathbf{A}$ but not at $\mathbf{A}^*$. Furthermore, you can't replace $(\star)^T$ by $(\star)^*$ in the above argument because if $p(z)$ has some complex coefficients, then $\big[p(\mathbf{A})\big]^* \neq p(\mathbf{A}^*)$.

7.9.13. (a) If $f_1(z) = \mathrm{e}^z$, $f_2(z) = \mathrm{e}^{-z}$, and $p(x, y) = xy - 1$, then

$$h(z) = p\big(f_1(z),\, f_2(z)\big) = \mathrm{e}^z \mathrm{e}^{-z} - 1 = 0 \quad \text{for all} \quad z \in \mathcal{C},$$

so $h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big) = \mathbf{0}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$, and thus $\mathrm{e}^{\mathbf{A}}\mathrm{e}^{-\mathbf{A}} - \mathbf{I} = \mathbf{0}$.

(b) Use $f_1(z) = \mathrm{e}^{\alpha z}$, $f_2(z) = \big(\mathrm{e}^z\big)^\alpha$, and $p(x, y) = x - y$. Since

$$h(z) = p\big(f_1(z),\, f_2(z)\big) = \mathrm{e}^{\alpha z} - \big(\mathrm{e}^z\big)^\alpha = 0 \quad \text{for all } z \in \mathcal{C},$$

$h(\mathbf{A}) = p\big(f_1(\mathbf{A}),\, f_2(\mathbf{A})\big) = \mathbf{0}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$, and thus $\mathrm{e}^{\alpha\mathbf{A}} = \big(\mathrm{e}^{\mathbf{A}}\big)^\alpha$.

(c) Using $f_1(z) = \mathrm{e}^{\mathrm{i}z}$, $f_2(z) = \cos z + \mathrm{i}\sin z$, and $p(x,y) = x - y$ produces $h(z) = p\big(f_1(z),\, f_2(z)\big)$, which is zero for all z, so $h(\mathbf{A}) = \mathbf{0}$ for all $\mathbf{A} \in \mathcal{C}^{n\times n}$.

7.9.14. (a) The representation $\mathrm{e}^z = \sum_{n=0}^{\infty} z^n/n!$ together with $\mathbf{AB} = \mathbf{BA}$ yields

$$\mathrm{e}^{\mathbf{A}+\mathbf{B}} = \sum_{n=0}^{\infty} \frac{(\mathbf{A}+\mathbf{B})^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} \mathbf{A}^j \mathbf{B}^{n-j} = \sum_{n=0}^{\infty} \sum_{j=0}^{n} \frac{\mathbf{A}^j \mathbf{B}^{n-j}}{j!(n-j)!}$$
$$= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{1}{r!} \frac{1}{s!} \mathbf{A}^r \mathbf{B}^s = \left(\sum_{r=0}^{\infty} \frac{\mathbf{A}^r}{r!} \right) \left(\sum_{s=0}^{\infty} \frac{\mathbf{B}^s}{s!} \right) = \mathrm{e}^{\mathbf{A}} \mathrm{e}^{\mathbf{B}}.$$

(b) If $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, then

$$\mathrm{e}^{\mathbf{A}} \mathrm{e}^{\mathbf{B}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad \text{but} \quad \mathrm{e}^{\mathbf{A}+\mathbf{B}} = \frac{1}{2} \begin{pmatrix} \mathrm{e} + \mathrm{e}^{-1} & \mathrm{e} - \mathrm{e}^{-1} \\ \mathrm{e} - \mathrm{e}^{-1} & \mathrm{e} + \mathrm{e}^{-1} \end{pmatrix}.$$

7.9.15. The characteristic equation for $\mathbf{A}$ is $\lambda^3 = 0$, and the number of 2×2 Jordan blocks associated with $\lambda = 0$ is $\nu_2 = rank\,(\mathbf{A}) - 2\,rank\,\left(\mathbf{A}^2\right) + rank\,\left(\mathbf{A}^3\right) = 1$ (from the formula on p. 590), so $index(\lambda = 0) = 2$. Therefore, for $f(z) = \mathrm{e}^z$ we are looking for a polynomial $p(z) = \alpha_0 + \alpha_1 z$ such that $p(0) = f(0) = 1$ and $p'(0) = f'(0) = 1$. This yields the Hermite interpolation polynomial as

$$p(z) = 1 + z, \quad \text{so} \quad \mathrm{e}^{\mathbf{A}} = p(\mathbf{A}) = \mathbf{I} + \mathbf{A}.$$

Note: Since $\mathbf{A}^2 = 0$, this agrees with the infinite series representation for $\mathrm{e}^{\mathbf{A}}$.

7.9.16. (a) The advantage is that the only the algebraic multiplicity and not the index of each eigenvalue is required—determining index generally requires more effort. The disadvantage is that a higher-degree polynomial might be required, so a larger system might have to be solved. Another disadvantage is the fact that f may not have enough derivatives defined at some eigenvalue for this method to work in spite of the fact that $f(\mathbf{A})$ exists.

(b) The characteristic equation for $\mathbf{A}$ is $\lambda^3 = 0$, so, for $f(z) = \mathrm{e}^z$, we are looking for a polynomial $p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2$ such that $p(0) = f(0) = 1$, $p'(0) = f'(0) = 1$, and $p''(0) = f''(0) = 1$. This yields

$$p(z) = 1 + z + \frac{z^2}{2}, \quad \text{so} \quad \mathrm{e}^{\mathbf{A}} = p(\mathbf{A}) = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2}.$$

Note: Since $\mathbf{A}^2 = 0$, this agrees with the result in Exercise 7.9.15.

7.9.17. Since $\sigma\,(\mathbf{A}) = \{\alpha\}$ with $index(\alpha) = 3$, it follows from (7.9.9) that

$$f(\mathbf{A}) = f(\alpha)\mathbf{G}_1 + f'(\alpha)(\mathbf{A} - \alpha\mathbf{I})\mathbf{G}_1 + \frac{f''(\alpha)}{2!}(\mathbf{A} - \alpha\mathbf{I})^2\mathbf{G}_1.$$

The desired result is produced by using $\mathbf{G}_1 = \mathbf{I}$ (because of (7.9.10)), and $\mathbf{A} - \alpha\mathbf{I} = \beta\mathbf{N} + \gamma\mathbf{N}^2$, and $\mathbf{N}^3 = \mathbf{0}$.

7.9.18. Since

$$g(\mathbf{J}_\star) = \begin{pmatrix} g(\lambda) & g'(\lambda) & g''(\lambda)/2! \\ 0 & g(\lambda) & g'(\lambda) \\ 0 & 0 & g(\lambda) \end{pmatrix},$$

using Exercise 7.9.17 with $\alpha = g(\lambda)$, $\beta = g'(\lambda)$, and $\gamma = g''(\lambda)/2!$ yields

$$f\big(g(\mathbf{J}_\star)\big) = f(g(\lambda))\mathbf{I} + g'(\lambda)f'(g(\lambda))\mathbf{N} + \left[\frac{g''(\lambda)f'(g(\lambda))}{2!} + \frac{\big[g'(\lambda)\big]^2 f''(g(\lambda))}{2!}\right]\mathbf{N}^2.$$

Observing that

$$h(\mathbf{J}_\star) = \begin{pmatrix} h(\lambda) & h'(\lambda) & h''(\lambda)/2! \\ 0 & h(\lambda) & h'(\lambda) \\ 0 & 0 & h(\lambda) \end{pmatrix} = h(\lambda)\mathbf{I} + h'(\lambda)\mathbf{N} + \frac{h''(\lambda)}{2!}\mathbf{N}^2,$$

where $h(\lambda) = f(g(\lambda))$, $h'(\lambda) = f'(g(\lambda))g'(\lambda)$, and

$$h''(\lambda) = g''(\lambda)f'(g(\lambda)) + f''(g(\lambda))\big[g'(\lambda)\big]^2$$

proves that $h(\mathbf{J}_\star) = f\big(g(\mathbf{J}_\star)\big)$.

7.9.19. For the function

$$f_i(z) = \begin{cases} 1 & \text{in a small circle about } \lambda_i \text{ that is interior to } \Gamma_i, \\ 0 & \text{elsewhere,} \end{cases}$$

it follows, just as in Exercise 7.9.2, that $f_i(\mathbf{A}) = \mathbf{G}_i$. But using f_i in (7.9.22) produces $f_i(\mathbf{A}) = \frac{1}{2\pi\mathrm{i}}\int_{\Gamma_i}(\xi\mathbf{I} - \mathbf{A})^{-1}d\xi$, and thus the result is proven.

7.9.20. For a $k \times k$ Jordan block $\mathbf{J}_\star = \begin{pmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & \lambda \end{pmatrix}$, it's straightforward to verify that

$$\mathbf{J}_\star^{-1} = \begin{pmatrix} \lambda^{-1} & -\lambda^{-2} & \lambda^{-3} & \cdots & -1^{(k-1)}\lambda^{-k} \\ & \lambda^{-1} & -\lambda^{-2} & \ddots & \vdots \\ & & \ddots & \ddots & \lambda^{-3} \\ & & & \lambda^{-1} & -\lambda^{-2} \\ & & & & \lambda^{-1} \end{pmatrix} = \begin{pmatrix} f(\lambda) & f'(\lambda) & \dfrac{f''(\lambda)}{2!} & \cdots & \dfrac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & f'(\lambda) & \ddots & \vdots \\ & & \ddots & \ddots & \dfrac{f''(\lambda)}{2!} \\ & & & f(\lambda) & f'(\lambda) \\ & & & & f(\lambda) \end{pmatrix}$$

for $f(z) = z^{-1}$, and thus if $\mathbf{J} = \begin{pmatrix} \ddots & & \\ & \mathbf{J}_\star & \\ & & \ddots \end{pmatrix}$ is the Jordan form for $\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$, then the representation of $\mathbf{A}^{-1}$ as $\mathbf{A}^{-1} = \mathbf{P}\mathbf{J}^{-1}\mathbf{P}^{-1}$ agrees with the expression for $f(\mathbf{A}) = \mathbf{P}f(\mathbf{J})\mathbf{P}^{-1}$ given in (7.9.3) when $f(z) = z^{-1}$.

7.9.21. $\dfrac{1}{2\pi \mathrm{i}}\displaystyle\int_\Gamma \xi^{-1}(\xi\mathbf{I}-\mathbf{A})^{-1}d\xi = \mathbf{A}^{-1}.$

7.9.22. Partition the Jordan form for $\mathbf{A}$ as $\mathbf{J} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}$ in which $\mathbf{C}$ contains all Jordan segments associated with nonzero eigenvalues and $\mathbf{N}$ contains all Jordan segments associated with the zero eigenvalue (if one exists). Observe that $\mathbf{N}$ is nilpotent, so $g(\mathbf{N}) = \mathbf{0}$, and consequently

$$\mathbf{A} = \mathbf{P}\begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{pmatrix}\mathbf{P}^{-1} \Rightarrow g(\mathbf{A}) = \mathbf{P}\begin{pmatrix} g(\mathbf{C}) & \mathbf{0} \\ \mathbf{0} & g(\mathbf{N}) \end{pmatrix}\mathbf{P}^{-1} = \mathbf{P}\begin{pmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{P}^{-1} = \mathbf{A}^D.$$

It follows from Exercise 5.12.17 (p. 428) that $g(\mathbf{A})$ is the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ if and only if $\mathbf{A}$ is an RPN matrix.

7.9.23. Use the Cauchy–Goursat theorem to observe that $\int_\Gamma \xi^{-j}d\xi = 0$ for $j = 2, 3, \ldots,$ and follow the argument given in Example 7.9.8 (p. 611) with $\lambda_1 = 0$ along with the result of Exercise 7.9.22 to write

$$\begin{aligned}
\frac{1}{2\pi \mathrm{i}}\int_\Gamma \xi^{-1}(\xi\mathbf{I}-\mathbf{A})^{-1}d\xi &= \frac{1}{2\pi \mathrm{i}}\int_\Gamma \xi^{-1}\mathbf{R}(\xi)d\xi \\
&= \frac{1}{2\pi \mathrm{i}}\int_\Gamma \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{\xi^{-1}}{(\xi-\lambda_i)^{j+1}}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i d\xi \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\left[\frac{1}{2\pi \mathrm{i}}\int_\Gamma \frac{\xi^{-1}}{(\xi-\lambda_i)^{j+1}}d\xi\right](\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i \\
&= \sum_{i=1}^{s}\sum_{j=0}^{k_i-1}\frac{g^{(j)}(\lambda_i)}{j!}(\mathbf{A}-\lambda_i\mathbf{I})^j\mathbf{G}_i = g(\mathbf{A}) = \mathbf{A}^D.
\end{aligned}$$

Solutions for exercises in section 7. 10

7.10.1. The characteristic equation for $\mathbf{A}$ is $0 = x^3-(3/4)x-(1/4) = (x-1)(x-1/2)^2$, so (7.10.33) guarantees that $\mathbf{A}$ is convergent (and hence also summable).

The characteristic equation for $\mathbf{B}$ is $x^3-1 = 0$, so the eigenvalues are the three cube roots of unity, and thus (7.10.33) insures $\mathbf{B}$ is not convergent, but $\mathbf{B}$ is summable because $\rho(\mathbf{B}) = 1$ and each eigenvalue on the unit circle is semisimple (in fact, each eigenvalue is simple).

The characteristic equation for $\mathbf{C}$ is

$$0 = x^3 - (5/2)x^2 + 2x - (1/2) = (x-1)^2(x-1/2),$$

but $\mathit{index}(\lambda = 1) = 2$ because $\mathit{rank}\,(\mathbf{C}-\mathbf{I}) = 2$ while $1 = \mathit{rank}\,(\mathbf{C}-\mathbf{I})^2 = \mathit{rank}\,(\mathbf{C}-\mathbf{I})^3 = \cdots$, so $\mathbf{C}$ is neither convergent nor summable.

7.10.2. Since $\mathbf{A}$ is convergent, (7.10.41) says that a full-rank factorization $\mathbf{I}-\mathbf{A}=\mathbf{BC}$ can be used to compute $\lim_{k\to\infty}\mathbf{A}^k = \mathbf{G} = \mathbf{I}-\mathbf{B}(\mathbf{CB})^{-1}\mathbf{C}$. One full-rank factorization is obtained by placing the basic columns of $\mathbf{I}-\mathbf{A}$ in $\mathbf{B}$ and the nonzero rows of $\mathbf{E}_{(\mathbf{I}-\mathbf{A})}$ in $\mathbf{C}$. This yields

$$\mathbf{B}=\begin{pmatrix}-3/2 & -3/2\\ 1 & -1/2\\ 1 & -1/2\end{pmatrix},\quad \mathbf{C}=\begin{pmatrix}1 & -1 & 0\\ 0 & 0 & 1\end{pmatrix},\quad \text{and}\quad \mathbf{G}=\begin{pmatrix}0 & 1 & -1\\ 0 & 1 & -1\\ 0 & 0 & 0\end{pmatrix}.$$

Alternately, since $\lambda=1$ is a simple eigenvalue, the limit $\mathbf{G}$ can also be determined by computing right- and left-hand eigenvectors, $\mathbf{x}=(1,1,0)^T$ and $\mathbf{y}^T=(0,-1,1)$, associated with $\lambda=1$ and setting $\mathbf{G}=\mathbf{xy}^T/(\mathbf{y}^T\mathbf{x})$ as described in (7.2.12) on p. 518. The matrix $\mathbf{B}$ is not convergent but it is summable, and since the unit eigenvalue is simple, the Cesàro limit $\mathbf{G}$ can be determined as described in (7.2.12) on p. 518 by computing right- and left-hand eigenvectors, $\mathbf{x}=(1,1,1)^T$ and $\mathbf{y}^T=(1,1,1)$, associated with $\lambda=1$ and setting $\mathbf{G}=\mathbf{xy}^T/(\mathbf{y}^T\mathbf{x})=\dfrac{1}{3}\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix}$.

7.10.3. To see that $\mathbf{x}(k)=\mathbf{A}^k\mathbf{x}(0)$ solves $\mathbf{x}(k+1)=\mathbf{Ax}(k)$, use successive substitution to write $\mathbf{x}(1)=\mathbf{Ax}(0)$, $\mathbf{x}(2)=\mathbf{Ax}(1)=\mathbf{A}^2\mathbf{x}(0)$, $\mathbf{x}(3)=\mathbf{Ax}(2)=\mathbf{A}^3\mathbf{x}(0)$, etc. Of course you could build a formal induction argument, but it's not necessary. To see that $\mathbf{x}(k)=\mathbf{A}^k\mathbf{x}(0)+\sum_{j=0}^{k-1}\mathbf{A}^{k-j-1}\mathbf{b}(j)$ solves the nonhomogeneous equation $\mathbf{x}(k+1)=\mathbf{Ax}(k)+\mathbf{b}(k)$, use successive substitution to write

$$\begin{aligned}
&\mathbf{x}(1)=\mathbf{Ax}(0)+\mathbf{b}(0),\\
&\mathbf{x}(2)=\mathbf{Ax}(1)+\mathbf{b}(1)=\mathbf{A}^2\mathbf{x}(0)+\mathbf{Ab}(0)+\mathbf{b}(0),\\
&\mathbf{x}(3)=\mathbf{Ax}(2)+\mathbf{b}(2)=\mathbf{A}^3\mathbf{x}(0)+\mathbf{A}^2\mathbf{b}(0)+\mathbf{Ab}(0)+\mathbf{b}(0),\\
&\text{etc.}
\end{aligned}$$

7.10.4. Put the basic columns of $\mathbf{I}-\mathbf{A}$ in $\mathbf{B}$ and the nonzero rows of the reduced row echelon form $\mathbf{E}_{(\mathbf{I}-\mathbf{A})}$ in $\mathbf{C}$ to build a full-rank factorization of $\mathbf{I}-\mathbf{A}=\mathbf{BC}$ (Exercise 3.9.8, p. 140), and use (7.10.41).

$$\mathbf{p}=\mathbf{Gp}(0)=(\mathbf{I}-\mathbf{B}(\mathbf{CB})^{-1}\mathbf{C})\mathbf{p}(0)=\begin{pmatrix}1/6 & 1/6 & 1/6 & 1/6\\ 1/3 & 1/3 & 1/3 & 1/3\\ 1/3 & 1/3 & 1/3 & 1/3\\ 1/6 & 1/6 & 1/6 & 1/6\end{pmatrix}\begin{pmatrix}p_1(0)\\ p_2(0)\\ p_3(0)\\ p_4(0)\end{pmatrix}=\begin{pmatrix}1/6\\ 1/3\\ 1/3\\ 1/6\end{pmatrix}.$$

7.10.5. To see that $\mathbf{x}(k)=\mathbf{A}^k\mathbf{x}(0)$ solves $\mathbf{x}(k+1)=\mathbf{Ax}(k)$, use successive substitution to write $\mathbf{x}(1)=\mathbf{Ax}(0)$, $\mathbf{x}(2)=\mathbf{Ax}(1)=\mathbf{A}^2\mathbf{x}(0)$, $\mathbf{x}(3)=\mathbf{Ax}(2)=\mathbf{A}^3\mathbf{x}(0)$, etc. Of course you could build a formal induction argument, but it's not necessary.

To see that $\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) + \sum_{j=0}^{k-1} \mathbf{A}^{k-j-1}\mathbf{b}(j)$ solves the nonhomogeneous equation $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}(k)$, use successive substitution to write

$$\begin{aligned}
&\mathbf{x}(1) = \mathbf{A}\mathbf{x}(0) + \mathbf{b}(0),\\
&\mathbf{x}(2) = \mathbf{A}\mathbf{x}(1) + \mathbf{b}(1) = \mathbf{A}^2\mathbf{x}(0) + \mathbf{A}\mathbf{b}(0) + \mathbf{b}(0),\\
&\mathbf{x}(3) = \mathbf{A}\mathbf{x}(2) + \mathbf{b}(2) = \mathbf{A}^3\mathbf{x}(0) + \mathbf{A}^2\mathbf{b}(0) + \mathbf{A}\mathbf{b}(0) + \mathbf{b}(0),\\
&\text{etc.}
\end{aligned}$$

7.10.6. Use (7.1.12) on p. 497 along with (7.10.5) on p. 617.

7.10.7. For $\mathbf{A}_1$, the respective iteration matrices for Jacobi and Gauss–Seidel are

$$\mathbf{H}_J = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{H}_{GS} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}.$$

$\mathbf{H}_J$ is nilpotent of index three, so $\sigma(\mathbf{H}_J) = \{0\}$, and hence $\rho(\mathbf{H}_J) = 0 < 1$. Clearly, $\mathbf{H}_{GS}$ is triangular, so $\rho(\mathbf{H}_{GS}) = 2. > 1$ Therefore, for arbitrary right-hand sides, Jacobi's method converges after two steps, whereas the Gauss–Seidel method diverges. On the other hand, for $\mathbf{A}_2$,

$$\mathbf{H}_J = \frac{1}{2}\begin{pmatrix} 0 & 1 & -1 \\ -2 & 0 & -2 \\ 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{H}_{GS} = \frac{1}{2}\begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{pmatrix},$$

and a little computation reveals that $\sigma(\mathbf{H}_J) = \left\{\pm i\sqrt{5}/2\right\}$, so $\rho(\mathbf{H}_J) > 1$, while $\rho(\mathbf{H}_{GS}) = 1/2 < 1$. These examples show that there is no universal superiority enjoyed by either method because there is no universal domination of $\rho(\mathbf{H}_J)$ by $\rho(\mathbf{H}_{GS})$, or vise versa.

7.10.8. (a) $\det(\alpha\mathbf{D} - \mathbf{L} - \mathbf{U}) = \det\left(\alpha\mathbf{D} - \beta\mathbf{L} - \beta^{-1}\mathbf{U}\right) = 8\alpha^3 - 4\alpha$ for all real α and $\beta \neq 0$. Furthermore, the Jacobi iteration matrix is

$$\mathbf{H}_J = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix},$$

and Example 7.2.5, p. 514, shows $\sigma(\mathbf{H}_J) = \{\cos(\pi/4), \cos(2\pi/4), \cos(3\pi/4)\}$. Clearly, these eigenvalues are real and $\rho(\mathbf{H}_J) = (1/\sqrt{2}) \approx .707 < 1$.

(b) According to (7.10.24),

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{H}_J)}} \approx 1.172, \qquad \text{and} \qquad \rho\left(\mathbf{H}_{\omega_{\text{opt}}}\right) = \omega_{\text{opt}} - 1 \approx .172.$$

(c) $R_J = -\log_{10}\rho(\mathbf{H}_J) = \log_{10}(\sqrt{2}) \approx .1505$, $R_{GS} = 2R_J \approx .301$, and $R_{\text{opt}} = -\log_{10}\rho(\mathbf{H}_{\text{opt}}) \approx .766$.

(d) I used standard IEEE 64-bit floating-point arithmetic (i.e., about 16 decimal digits of precision) for all computations, but I rounded the results to 3 places to report the answers given below. Depending on your own implementation, your answers may vary slightly.

Jacobi with 21 iterations:

1	1.5	2.5	3.25	3.75	4.12	4.37	4.56	4.69	4.78	4.84	4.89	4.92	4.95	4.96	4.97	4.98	4.99	4.99	4.99	5	5
1	3	4.5	5.5	6.25	6.75	7.12	7.37	7.56	7.69	7.78	7.84	7.89	7.92	7.95	7.96	7.97	7.98	7.99	7.99	7.99	8
1	3.5	4.5	5.25	5.75	6.12	6.37	6.56	6.69	6.78	6.84	6.89	6.92	6.95	6.96	6.97	6.98	6.99	6.99	6.99	7	7

Gauss–Seidel with 11 iterations:

1	1.5	2.62	3.81	4.41	4.7	4.85	4.93	4.96	4.98	4.99	5
1	3.25	5.62	6.81	7.41	7.7	7.85	7.93	7.96	7.98	7.99	8
1	4.62	5.81	6.41	6.7	6.85	6.93	6.96	6.98	6.99	7	7

SOR (optimum) with 6 iterations:

1	1.59	3.06	4.59	4.89	4.98	5
1	3.69	6.73	7.69	7.93	7.99	8
1	5.5	6.51	6.9	6.98	7	7

7.10.9. The product rule for determinants produces

$$\det(\mathbf{H}_\omega) = \det\left[(\mathbf{D}-\omega\mathbf{L})^{-1}\right]\det\left[(1-\omega)\mathbf{D}+\omega\mathbf{U}\right] = \prod_{i=1}^{n} a_{ii}^{-1}\prod_{i=1}^{n}(1-\omega)a_{ii} = (1-\omega)^n.$$

But it's also true that $\det(\mathbf{H}_\omega) = \prod_{i=1}^{n}\lambda_i$, where the λ_i's are the eigenvalues of $\mathbf{H}_\omega$. Consequently, $|\lambda_k| \geq |1-\omega|$ for some k because if $|\lambda_i| < |1-\omega|$ for all i, then $|1-\omega|^n = |\det(\mathbf{H}_\omega)| = \prod_i |\lambda_i| < |1+\omega|^n$, which is impossible. Therefore, $|1-\omega| \leq |\lambda_k| \leq \rho(\mathbf{H}_\omega) < 1$ implies $0 < \omega < 2$.

7.10.10. Observe that $\mathbf{H}_J$ is the block-triangular matrix

$$\mathbf{H}_J = \frac{1}{4}\begin{pmatrix} \mathbf{K} & \mathbf{I} & & & \\ \mathbf{I} & \mathbf{K} & \mathbf{I} & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{I} & \mathbf{K} & \mathbf{I} \\ & & & \mathbf{I} & \mathbf{K} \end{pmatrix}_{n^2\times n^2} \quad \text{with} \quad \mathbf{K} = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{pmatrix}_{n\times n}.$$

Proceed along the same lines as in Example 7.6.2, to argue that $\mathbf{H}_J$ is similar to the block-diagonal matrix

$$\begin{pmatrix} \mathbf{T}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{T}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{T}_n \end{pmatrix}, \quad \text{where} \quad \mathbf{T}_i = \begin{pmatrix} \kappa_i & 1 & & & \\ 1 & \kappa_i & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & \kappa_i & 1 \\ & & & 1 & \kappa_i \end{pmatrix}_{n\times n}$$

in which $\kappa_i \in \sigma(\mathbf{K})$. Use the result of Example 7.2.5 (p. 514) to infer that the eigenvalues of $\mathbf{T}_i$ are $\kappa_i + 2\cos j\pi/(n+1)$ for $j = 1, 2, \ldots, n$ and, similarly, the eigenvalues of $\mathbf{K}$ are $\kappa_i = 2\cos i\pi/(n+1)$ for $i = 1, 2, \ldots, n$. Consequently the n^2 eigenvalues of $\mathbf{H}_J$ are $\lambda_{ij} = (1/4)\big[2\cos i\pi/(n+1) + 2\cos j\pi/(n+1)\big]$, so $\rho(\mathbf{H}_J) = \max_{i,j} \lambda_{ij} = \cos\pi/(n+1)$.

7.10.11. If $\lim_{n\to\infty} \alpha_n = \alpha$, then for each $\epsilon > 0$ there is a natural number $N = N(\epsilon)$ such that $|\alpha_n - \alpha| < \epsilon/2$ for all $n \geq N$. Furthermore, there exists a real number β such that $|\alpha_n - \alpha| < \beta$ for all n. Consequently, for all $n \geq N$,

$$|\mu_n - \alpha| = \left|\frac{\alpha_1 + \alpha_2 + \cdots + \alpha_n}{n} - \alpha\right| = \frac{1}{n}\left|\sum_{k=1}^{N}(\alpha_k - \alpha) + \sum_{k=N+1}^{n}(\alpha_k - \alpha)\right|$$

$$\leq \frac{1}{n}\sum_{k=1}^{N}|\alpha_k - \alpha| + \frac{1}{n}\sum_{k=N+1}^{n}|\alpha_k - \alpha| < \frac{N\beta}{n} + \frac{n-N}{n}\frac{\epsilon}{2} \leq \frac{N\beta}{n} + \frac{\epsilon}{2}.$$

When n is sufficiently large, $N\beta/n \leq \epsilon/2$ so that $|\mu_n - \alpha| < \epsilon$, and therefore, $\lim_{n\to\infty} \mu_n = \alpha$. **Note:** The same proof works for vectors and matrices by replacing $|\star|$ with a vector or matrix norm.

7.10.12. Prove that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (e) $\Rightarrow$ (f) $\Rightarrow$ (a).

(a) $\Rightarrow$ (b): This is a consequence of (7.10.28).

(b) $\Rightarrow$ (c): Use induction on the size of $\mathbf{A}_{n\times n}$. For $n = 1$, the result is trivial. Suppose the result holds for $n = k$—i.e., suppose positive leading minors insures the existence of LU factors which are M-matrices when $n = k$. For $n = k+1$, use the induction hypothesis to write

$$\mathbf{A}_{(k+1)\times(k+1)} = \begin{pmatrix} \widetilde{\mathbf{A}} & \mathbf{c} \\ \mathbf{d}^T & \alpha \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{L}}\widetilde{\mathbf{U}} & \mathbf{c} \\ \mathbf{d}^T & \alpha \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{L}} & \mathbf{0} \\ \mathbf{d}^T\widetilde{\mathbf{U}}^{-1} & 1 \end{pmatrix}\begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{L}}^{-1}\mathbf{c} \\ \mathbf{0} & \sigma \end{pmatrix} = \mathbf{LU},$$

where $\widetilde{\mathbf{L}}$ and $\widetilde{\mathbf{U}}$ are M-matrices. Notice that $\sigma > 0$ because $\det(\widetilde{\mathbf{U}}) > 0$ and $0 < \det(\mathbf{A}) = \sigma\det(\widetilde{\mathbf{L}})\det(\widetilde{\mathbf{U}})$. Consequently, $\mathbf{L}$ and $\mathbf{U}$ are M-matrices because

$$\mathbf{L}^{-1} = \begin{pmatrix} \widetilde{\mathbf{L}}^{-1} & \mathbf{0} \\ -\mathbf{d}^T\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{L}}^{-1} & 1 \end{pmatrix} \geq \mathbf{0} \quad \text{and} \quad \mathbf{U}^{-1} = \begin{pmatrix} \widetilde{\mathbf{U}}^{-1} & -\sigma^{-1}\widetilde{\mathbf{U}}^{-1}\widetilde{\mathbf{L}}^{-1}\mathbf{c} \\ \mathbf{0} & \sigma^{-1} \end{pmatrix} \geq \mathbf{0}.$$

(c) $\Rightarrow$ (d): $\mathbf{A} = \mathbf{LU}$ with $\mathbf{L}$ and $\mathbf{U}$ M-matrices implies $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1} \geq \mathbf{0}$, so if $\mathbf{x} = \mathbf{A}^{-1}\mathbf{e}$, where $\mathbf{e} = (1, 1, \ldots, 1)^T$, then $\mathbf{x} > \mathbf{0}$ (otherwise $\mathbf{A}^{-1}$ would have a zero row, which would force $\mathbf{A}$ to be singular), and $\mathbf{Ax} = \mathbf{e} > \mathbf{0}$.

(d) $\Rightarrow$ (e): If $\mathbf{x} > \mathbf{0}$ is such that $\mathbf{Ax} > \mathbf{0}$, define $\mathbf{D} = \operatorname{diag}(x_1, x_2, \ldots, x_n)$ and set $\mathbf{B} = \mathbf{AD}$, which is clearly another Z-matrix. For $\mathbf{e} = (1, 1, \ldots, 1)^T$, notice that $\mathbf{Be} = \mathbf{ADe} = \mathbf{Ax} > \mathbf{0}$ says each row sum of $\mathbf{B} = \mathbf{AD}$ is positive. In other words, for each $i = 1, 2, \ldots, n$,

$$0 < \sum_j b_{ij} = \sum_{j\neq i} b_{ij} + b_{ii} \;\Rightarrow\; b_{ii} > \sum_{j\neq i} -b_{ij} = \sum_{j\neq i}|b_{ij}| \quad \text{for each } i = 1, 2, \ldots, n.$$

(e) $\Rightarrow$ (f): Suppose that $\mathbf{AD}$ is diagonally dominant for a diagonal matrix $\mathbf{D}$ with positive entries, and suppose each $a_{ii} > 0$. If $\mathbf{E} = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ and $-\mathbf{N}$ is the matrix containing the off-diagonal entries of $\mathbf{A}$, then $\mathbf{A} = \mathbf{E} - \mathbf{N}$ is the Jacobi splitting for $\mathbf{A}$ as described in Example 7.10.4 on p. 622, and $\mathbf{AD} = \mathbf{ED} + \mathbf{ND}$ is the Jacobi splitting for $\mathbf{AD}$ with the iteration matrix $\mathbf{H} = \mathbf{D}^{-1}\mathbf{E}^{-1}\mathbf{ND}$. It was shown in Example 7.10.4 that diagonal dominance insures convergence of Jacobi's method (i.e., $\rho(\mathbf{H}) < 1$), so, by (7.10.14), p. 620,

$$\mathbf{A} = \mathbf{ED}(\mathbf{I} - \mathbf{H})\mathbf{D}^{-1} \quad \Longrightarrow \quad \mathbf{A}^{-1} = \mathbf{D}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{D}^{-1}\mathbf{E}^{-1} \geq \mathbf{0},$$

and this guarantees that if $\mathbf{Ax} \geq \mathbf{0}$, then $\mathbf{x} \geq \mathbf{0}$.

(f) $\Rightarrow$ (a): Let $r \geq \max |a_{ii}|$ so that $\mathbf{B} = r\mathbf{I} - \mathbf{A} \geq \mathbf{0}$, and first show that the condition $(\mathbf{Ax} \geq \mathbf{0} \Rightarrow \mathbf{x} \geq \mathbf{0})$ insures the existence of $\mathbf{A}^{-1}$. For any $\mathbf{x} \in N(\mathbf{A})$,

$$(r\mathbf{I} - \mathbf{B})\mathbf{x} = \mathbf{0} \Rightarrow r\mathbf{x} = \mathbf{Bx} \Rightarrow r|\mathbf{x}| \leq |\mathbf{B}|\mathbf{x}| \Rightarrow \mathbf{A}(-|\mathbf{x}|) \geq \mathbf{0} \Rightarrow -|\mathbf{x}| \geq \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0},$$

so $N(\mathbf{A}) = \mathbf{0}$. Now, $\mathbf{A}[\mathbf{A}^{-1}]_{*i} = \mathbf{e}_i \geq \mathbf{0} \Rightarrow [\mathbf{A}^{-1}]_{*i} \geq \mathbf{0}$, and thus $\mathbf{A}^{-1} \geq \mathbf{0}$.

7.10.13. (a) If $\mathbf{M}_i$ is $n_i \times n_i$ with $rank(\mathbf{M}_i) = r_i$, then $\mathbf{B}_i$ is $n_i \times r_i$ and $\mathbf{C}_i$ is $r_i \times n_i$ with $rank(\mathbf{B}_i) = rank(\mathbf{C}_i) = r_i$. This means that $\mathbf{M}_{i+1} = \mathbf{C}_i\mathbf{B}_i$ is $r_i \times r_i$, so if $r_i < n_i$, then $\mathbf{M}_{i+1}$ has smaller size than $\mathbf{M}_i$. Since this can't happen indefinitely, there must be a point in the process at which $r_k = n_k$ or $r_k = 0$ and thus some $\mathbf{M}_k$ is either nonsingular or zero.

(b) Let $\mathbf{M} = \mathbf{M}_1 = \mathbf{A} - \lambda\mathbf{I}$, and notice that

$$\begin{aligned}
\mathbf{M}^2 &= \mathbf{B}_1\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1 = \mathbf{B}_1\mathbf{M}_2\mathbf{C}_1, \\
\mathbf{M}^3 &= \mathbf{B}_1\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1\mathbf{B}_1\mathbf{C}_1 = \mathbf{B}_1(\mathbf{B}_2\mathbf{C}_2)(\mathbf{B}_2\mathbf{C}_2)\mathbf{C}_1 = \mathbf{B}_1\mathbf{B}_2\mathbf{M}_3\mathbf{C}_2\mathbf{C}_1, \\
&\vdots \\
\mathbf{M}^i &= \mathbf{B}_1\mathbf{B}_2\cdots\mathbf{B}_{i-1}\mathbf{M}_i\mathbf{C}_{i-1}\cdots\mathbf{C}_2\mathbf{C}_1.
\end{aligned}$$

In general, it's true that $rank(\mathbf{XYZ}) = rank(\mathbf{Y})$ whenever $\mathbf{X}$ has full column rank and $\mathbf{Z}$ has full row rank (Exercise 4.5.12, p. 220), so applying this yields $rank(\mathbf{M}^i) = rank(\mathbf{M}_i)$ for each $i = 1, 2, \ldots$. Suppose that some $\mathbf{M}_i = \mathbf{C}_{i-1}\mathbf{B}_{i-1}$ is $n_i \times n_i$ and nonsingular. For this to happen, we must have $\mathbf{M}_{i-1} = \mathbf{B}_{i-1}\mathbf{C}_{i-1}$, where $\mathbf{B}_{i-1}$ is $n_{i-1} \times n_i$, $\mathbf{C}_{i-1}$ is $n_i \times n_{i-1}$, and

$$rank(\mathbf{M}_{i-1}) = rank(\mathbf{B}_{i-1}) = rank(\mathbf{C}_{i-1}) = n_i = rank(\mathbf{M}_i).$$

Therefore, if k is the smallest positive integer such that $\mathbf{M}_k^{-1}$ exists, then k is the smallest positive integer such that $rank(\mathbf{M}_{k-1}) = rank(\mathbf{M}_k)$, and thus k is the smallest positive integer such that $rank(\mathbf{M}^{k-1}) = rank(\mathbf{M}^k)$, which means that $index(\mathbf{M}) = k - 1$ or, equivalently, $index(\lambda) = k - 1$. On the other hand, if some $\mathbf{M}_i = \mathbf{0}$, then $rank(\mathbf{M}^i) = rank(\mathbf{M}_i)$ insures that $\mathbf{M}^i = \mathbf{0}$.

Consequently, if k is the smallest positive integer such that $\mathbf{M}_k = \mathbf{0}$, then k is the smallest positive integer such that $\mathbf{M}^k = \mathbf{0}$. Therefore, $\mathbf{M}$ is nilpotent of index k, and this implies that $index(\lambda) = k$.

7.10.14. $\mathbf{M} = \mathbf{A} - 4\mathbf{I} = \begin{pmatrix} -7 & -8 & -9 \\ 5 & 7 & 9 \\ -1 & -2 & -3 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow \mathbf{B}_1 = \begin{pmatrix} -7 & -8 \\ 5 & 7 \\ -1 & -2 \end{pmatrix}$ and $\mathbf{C}_1 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$, so $\mathbf{M}_2 = \mathbf{C}_1\mathbf{B}_1 = \begin{pmatrix} -6 & -6 \\ 3 & 3 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \Rightarrow$ $\mathbf{B}_2 = \begin{pmatrix} -6 \\ 3 \end{pmatrix}$ and $\mathbf{C}_2 = \begin{pmatrix} 1 & 1 \end{pmatrix}$, so $\mathbf{M}_3 = \mathbf{C}_2\mathbf{B}_2 = -3$. Since $\mathbf{M}_3$ is the first $\mathbf{M}_i$ to be nonsingular, $index(4) = 3 - 1 = 2$. Now, $index(1)$ if forced to be 1 because $1 = alg\ mult(1) \geq index(1) \geq 1$.

7.10.15. (a) Since $\sigma(\mathbf{A}) = \{1, 4\}$ with $index(1) = 1$ and $index(4) = 2$, the Jordan form for $\mathbf{A}$ is $\mathbf{J} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 1 \\ 0 & 0 & 4 \end{pmatrix}$.

(b) The Hermite interpolation polynomial $p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2$ is determined by solving $p(1) = f(1)$, $p(4) = f(4)$, and $p'(4) = f'(4)$ for α_i's. So

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 16 \\ 0 & 1 & 8 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} f(1) \\ f(4) \\ f'(4) \end{pmatrix} \implies \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 16 \\ 0 & 1 & 8 \end{pmatrix}^{-1} \begin{pmatrix} f(1) \\ f(4) \\ f'(4) \end{pmatrix}$$

$$= -\frac{1}{9} \begin{pmatrix} -16 & 7 & -12 \\ 8 & -8 & 15 \\ -1 & 1 & -3 \end{pmatrix} \begin{pmatrix} f(1) \\ f(4) \\ f'(4) \end{pmatrix}$$

$$= -\frac{1}{9} \begin{pmatrix} -16f(1) + 7f(4) - 12f'(4) \\ 8f(1) - 8f(4) + 15f'(4) \\ -f(1) + f(4) - 3f'(4) \end{pmatrix}.$$

Writing $f(\mathbf{A}) = p(\mathbf{A})$ produces

$$f(\mathbf{A}) = \left[\frac{-16\mathbf{I} + 8\mathbf{A} - \mathbf{A}^2}{-9}\right] f(1) + \left[\frac{7\mathbf{I} - 8\mathbf{A} + \mathbf{A}^2}{-9}\right] f(4) + \left[\frac{-12\mathbf{I} + 15\mathbf{A} - 3\mathbf{A}^2}{-9}\right] f'(4).$$

7.10.16. Suppose that $\lim_{k\to\infty} \mathbf{A}^k$ exists and is nonzero. It follows from (7.10.33) that $\lambda = 1$ is a semisimple eigenvalue of $\mathbf{A}$, so the Jordan form for $\mathbf{B}$ looks like $\mathbf{B} = \mathbf{I} - \mathbf{A} = \mathbf{P} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{K} \end{pmatrix} \mathbf{P}^{-1}$, where $\mathbf{I} - \mathbf{K}$ is nonsingular. Therefore, $\mathbf{B}$ belongs to a matrix group and

$$\mathbf{B}^{\#} = \mathbf{P} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} - \mathbf{K})^{-1} \end{pmatrix} \mathbf{P}^{-1} \implies \mathbf{I} - \mathbf{B}\mathbf{B}^{\#} = \mathbf{P} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{P}^{-1}.$$

Comparing $\mathbf{I} - \mathbf{B}\mathbf{B}^{\#}$ with (7.10.32) shows that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{I} - \mathbf{B}\mathbf{B}^{\#}$. If $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}$, then $\rho(\mathbf{A}) < 1$, and hence $\mathbf{B}$ is nonsingular, so $\mathbf{B}^{\#} = \mathbf{B}^{-1}$ and $\mathbf{I} - \mathbf{B}\mathbf{B}^{\#} = \mathbf{0}$. In other words, it's still true that $\lim_{k\to\infty} \mathbf{A}^k = \mathbf{I} - \mathbf{B}\mathbf{B}^{\#}$.

7.10.17. We already know from the development of (7.10.41) that if $rank\,(\mathbf{M}) = r$, then $\mathbf{CB}$ and $\mathbf{V}_1^*\mathbf{U}_1$ are $r \times r$ nonsingular matrices. It's a matter of simple algebra to verify that $\mathbf{M}\mathbf{M}^{\#}\mathbf{M} = \mathbf{M}$, $\mathbf{M}^{\#}\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}$, and $\mathbf{M}\mathbf{M}^{\#} = \mathbf{M}^{\#}\mathbf{M}$.

Solutions for exercises in section 7. 11

7.11.1. $m(x) = x^2 - 3x + 2$

7.11.2. $v(x) = x - 2$

7.11.3. $c(x) = (x-1)(x-2)^2$

7.11.4. $\mathbf{J} = \begin{pmatrix} \lambda & & & & & \\ & \lambda & & & & \\ & & \lambda & & & \\ & & & \mu & & \\ & & & & \mu & \\ & & & & \mu & 1 \\ & & & & & \mu \end{pmatrix}$

7.11.5. Set $\nu_0 = \|\mathbf{I}\|_F = 2$, $\mathbf{U}_0 = \mathbf{I}/2$, and generate the sequence (7.11.2).

$$\begin{aligned} r_{01} &= \langle \mathbf{U}_0 | \mathbf{A} \rangle = 2, \\ \nu_1 &= \|\mathbf{A} - r_{01}\mathbf{U}_0\|_F = \sqrt{1209}, \quad \mathbf{U}_1 = \frac{\mathbf{A} - r_{01}\mathbf{U}_0}{\nu_1} = \frac{\mathbf{A} - \mathbf{I}}{\sqrt{1209}}, \\ r_{02} &= \left\langle \mathbf{U}_0 \middle| \mathbf{A}^2 \right\rangle = 2, \quad r_{12} = \left\langle \mathbf{U}_1 \middle| \mathbf{A}^2 \right\rangle = 2\sqrt{1209}, \\ \nu_2 &= \|\mathbf{A}^2 - r_{02}\mathbf{U}_0 - r_{12}\mathbf{U}_1\|_F = 0, \end{aligned}$$

so that

$$\mathbf{R} = \begin{pmatrix} 2 & 2 \\ 0 & \sqrt{1209} \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 2 \\ 2\sqrt{1209} \end{pmatrix}, \text{ and } \mathbf{R}^{-1}\mathbf{c} = \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}.$$

Consequently, the minimum polynomial is $m(x) = x^2 - 2x + 1 = (x-1)^2$. As a by-product, we see that $\lambda = 1$ is the only eigenvalue of $\mathbf{A}$, and $index(\lambda) = 2$, so the Jordan form for $\mathbf{A}$ must be $\mathbf{J} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

7.11.6. Similar matrices have the same minimum polynomial because similar matrices have the same Jordan form, and hence they have the same eigenvalues with the same indicies.

7.11.10. $\mathbf{x} = (3, -1, -1)^T$

7.11.12. $\mathbf{x} = (-3, 6, 5)^T$

Solutions for Chapter 8

Solutions for exercises in section 8. 2

8.2.1. The eigenvalues are $\sigma(\mathbf{A}) = \{12, 6\}$ with $alg\ mult_{\mathbf{A}}(6) = 2$, and it's clear that $12 = \rho(\mathbf{A}) \in \sigma(\mathbf{A})$. The eigenspace $N(\mathbf{A} - 12\mathbf{I})$ is spanned by $\mathbf{e} = (1,1,1)^T$, so the Perron vector is $\mathbf{p} = (1/3)(1,1,1)^T$. The left-hand eigenspace $N(\mathbf{A}^T - 12\mathbf{I})$ is spanned by $(1,2,3)^T$, so the left-hand Perron vector is $\mathbf{q}^T = (1/6)(1,2,3)$.

8.2.3. If $\mathbf{p}_1$ and $\mathbf{p}_2$ are two vectors satisfying $\mathbf{A}\mathbf{p} = \rho(\mathbf{A})\mathbf{p}$, $\mathbf{p} > \mathbf{0}$, and $\|\mathbf{p}\|_1 = 1$, then $\dim N(\mathbf{A} - \rho(\mathbf{A})\mathbf{I}) = 1$ implies that $\mathbf{p}_1 = \alpha\mathbf{p}_2$ for some $\alpha < 0$. But $\|\mathbf{p}_1\|_1 = \|\mathbf{p}_2\|_1 = 1$ insures that $\alpha = 1$.

8.2.4. $\sigma(\mathbf{A}) = \{0, 1\}$, so $\rho(\mathbf{A}) = 1$ is the Perron root, and the Perron vector is $\mathbf{p} = (\alpha + \beta)^{-1}(\beta, \alpha)$.

8.2.5. (a) $\rho(\mathbf{A}/r) = 1$ is a simple eigenvalue of $\mathbf{A}/r$, and it's the only eigenvalue on the spectral circle of $\mathbf{A}/r$, so (7.10.33) on p. 630 guarantees that $\lim_{k\to\infty}(\mathbf{A}/r)^k$ exists.

(b) This follows from (7.10.34) on p. 630.

(c) $\mathbf{G}$ is the spectral projector associated with the simple eigenvalue $\lambda = r$, so formula (7.2.12) on p. 518 applies.

8.2.6. If $\mathbf{e}$ is the column of all 1 's, then $\mathbf{A}\mathbf{e} = \rho\mathbf{e}$. Since $\mathbf{e} > \mathbf{0}$, it must be a positive multiple of the Perron vector $\mathbf{p}$, and hence $\mathbf{p} = n^{-1}\mathbf{e}$. Therefore, $\mathbf{A}\mathbf{p} = \rho\mathbf{p}$ implies that $\rho = \rho(\mathbf{A})$. The result for column sums follows by considering $\mathbf{A}^T$.

8.2.7. Since $\rho = \max_i \sum_j a_{ij}$ is the largest row sum of $\mathbf{A}$, there must exist a matrix $\mathbf{E} \geq \mathbf{0}$ such that every row sum of $\mathbf{B} = \mathbf{A} + \mathbf{E}$ is ρ. Use Example 7.10.2 (p. 619) together with Exercise 8.2.7 to obtain $\rho(\mathbf{A}) \leq \rho(\mathbf{B}) = \rho$. The lower bound follows from the Collatz–Wielandt formula. If $\mathbf{e}$ is the column of ones, then $\mathbf{e} \in \mathcal{N}$, so

$$\rho(\mathbf{A}) = \max_{\mathbf{x}\in\mathcal{N}} f(\mathbf{x}) \geq f(\mathbf{e}) = \min_{1\leq i\leq n} \frac{[\mathbf{A}\mathbf{e}]_i}{e_i} = \min_i \sum_{j=1}^{n} a_{ij}.$$

8.2.8. (a), (b), (c), and (d) are illustrated by using the nilpotent matrix $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$.

(e) $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ has eigenvalues ± 1.

8.2.9. If $\xi = g(\mathbf{x})$ for $\mathbf{x} \in \mathcal{P}$, then $\xi\mathbf{x} \geq \mathbf{A}\mathbf{x} > \mathbf{0}$. Let $\mathbf{p}$ and $\mathbf{q}^T$ be the respective the right-hand and left-hand Perron vectors for $\mathbf{A}$ associated with the Perron root r, and use (8.2.3) along with $\mathbf{q}^T\mathbf{x} > 0$ to write

$$\xi\mathbf{x} \geq \mathbf{A}\mathbf{x} > \mathbf{0} \implies \xi\mathbf{q}^T\mathbf{x} \geq \mathbf{q}^T\mathbf{A}\mathbf{x} = r\mathbf{q}^T\mathbf{x} \implies \xi \geq r,$$

so $g(\mathbf{x}) \geq r$ for all $\mathbf{x} \in \mathcal{P}$. Since $g(\mathbf{p}) = r$ and $\mathbf{p} \in \mathcal{P}$, it follows that $r = \min_{\mathbf{x}\in\mathcal{P}} g(\mathbf{x})$.

8.2.10. $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \implies \rho(\mathbf{A}) = 5$, but $g(\mathbf{e}_1) = 1 \implies \min_{\mathbf{x}\in\mathcal{N}} g(\mathbf{x}) < \rho(\mathbf{A})$.

Solutions for exercises in section 8. 3

8.3.1. (a) The graph is strongly connected.

(b) $\rho(\mathbf{A}) = 3$, and $\mathbf{p} = (1/6,\, 1/2,\, 1/3)^T$.

(c) $h = 2$ because $\mathbf{A}$ is imprimitive and singular.

8.3.2. If $\mathbf{A}$ is nonsingular then there are either one or two distinct nonzero eigenvalues inside the spectral circle. But this is impossible because $\sigma(\mathbf{A})$ has to be invariant under rotations of $120°$ by the result on p. 677. Similarly, if $\mathbf{A}$ is singular with $alg\ mult_{\mathbf{A}}(0) = 1$, then there is a single nonzero eigenvalue inside the spectral circle, which is impossible.

8.3.3. No! The matrix $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ has $\rho(\mathbf{A}) = 2$ with a corresponding eigenvector $\mathbf{e} = (1, 1)^T$, but $\mathbf{A}$ is reducible.

8.3.4. $\mathbf{P}_n$ is nonnegative and irreducible (its graph is strongly connected), and $\mathbf{P}_n$ is imprimitive because $\mathbf{P}_n^n = \mathbf{I}$ insures that every power has zero entries. Furthermore, if $\lambda \in \sigma(\mathbf{P}_n)$, then $\lambda^n \in \sigma(\mathbf{P}_n^n) = \{1\}$, so all eigenvalues of $\mathbf{P}_n$ are roots of unity. Since all eigenvalues on the spectral circle are simple (recall (8.3.13) on p. 676) and uniformly distributed, it must be the case that $\sigma(\mathbf{P}_n) = \{1,\, \omega,\, \omega^2,\, \ldots,\, \omega^{n-1}\}$.

8.3.5. $\mathbf{A}$ is irreducible because the graph $\mathcal{G}(\mathbf{A})$ is strongly connected—every node is accessible by some sequence of paths from every other node.

8.3.6. $\mathbf{A}$ is imprimitive. This is easily seen by observing that each $\mathbf{A}^{2n}$ for $n > 1$ has the same zero pattern (and each $\mathbf{A}^{2n+1}$ for $n > 0$ has the same zero pattern), so every power of $\mathbf{A}$ has zero entries.

8.3.7. (a) Having row sums less than or equal to 1 means that $\|\mathbf{P}\|_\infty \leq 1$. Because $\rho(\star) \leq \|\star\|$ for every matrix norm (recall (7.1.12) on p. 497), it follows that $\rho(\mathbf{S}) \leq \|\mathbf{S}\|_1 \leq 1$.

(b) If $\mathbf{e}$ denotes the column of all 1's, then the hypothesis insures that $\mathbf{Se} \leq \mathbf{e}$, and $\mathbf{Se} \neq \mathbf{e}$. Since $\mathbf{S}$ is irreducible, the result in Example 8.3.1 (p. 674) implies that it's impossible to have $\rho(\mathbf{S}) = 1$ (otherwise $\mathbf{Se} = \mathbf{e}$), and therefore $\rho(\mathbf{S}) < 1$ by part (a).

8.3.8. If $\mathbf{p}$ is the Perron vector for $\mathbf{A}$, and if $\mathbf{e}$ is the column of 1's, then

$$\mathbf{D}^{-1}\mathbf{ADe} = \mathbf{D}^{-1}\mathbf{Ap} = r\mathbf{D}^{-1}\mathbf{p} = r\mathbf{e}$$

shows that every row sum of $\mathbf{D}^{-1}\mathbf{AD}$ is r, so we can take $\mathbf{P} = r^{-1}\mathbf{D}^{-1}\mathbf{AD}$ because the Perron–Frobenius theorem guarantees that $r > 0$.

8.3.9. Construct the Boolean matrices as described in Example 8.3.5 (p. 680), and show that $\mathbf{B}_9$ has a zero in the $(1, 1)$ position, but $\mathbf{B}_{10} > \mathbf{0}$.

8.3.10. According to the discussion on p. 630, $\mathbf{f}(t) \to \mathbf{0}$ if $r < 1$. If $r = 1$, then $\mathbf{f}(t) \to \mathbf{G}\mathbf{f}(0) = \mathbf{p}\Big(\mathbf{q}^T\mathbf{f}(0)/\mathbf{q}^T\mathbf{p}\Big) > \mathbf{0}$, and if $r > 1$, the results of the Leslie analysis imply that $f_k(t) \to \infty$ for each k.

8.3.11. The only nonzero coefficient in the characteristic equation for $\mathbf{L}$ is c_1, so $\gcd\{2, 3, \ldots, n\} = 1$.

8.3.12. (a) Suppose that $\mathbf{A}$ is essentially positive. Since we can always find a $\beta > 0$ such that $\beta\mathbf{I} + \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}) \geq \mathbf{0}$, and since $a_{ij} \geq 0$ for $i \neq j$, it follows that $\mathbf{A} + \beta\mathbf{I}$ is a nonnegative irreducible matrix, so (8.3.5) on p. 672 can be applied to conclude that $(\mathbf{A} + (1+\beta)\mathbf{I})^{n-1} > \mathbf{0}$, and thus $\mathbf{A} + \alpha\mathbf{I}$ is primitive with $\alpha = \beta + 1$. Conversely, if $\mathbf{A} + \alpha\mathbf{I}$ is primitive, then $\mathbf{A} + \alpha\mathbf{I}$ must be nonnegative and irreducible, and hence $a_{ij} \geq 0$ for every $i \neq j$, and $\mathbf{A}$ must be irreducible (diagonal entries don't affect the reducibility or irreducibility).

(b) If $\mathbf{A}$ is essentially positive, then $\mathbf{A} + \alpha\mathbf{I}$ is primitive for some α (by the first part), so $(\mathbf{A} + \alpha\mathbf{I})^k > \mathbf{0}$ for some k. Consequently, for all $t > 0$,

$$\mathbf{0} < \sum_{k=0}^{\infty} \frac{t^k(\mathbf{A} + \alpha\mathbf{I})^k}{k!} = \mathrm{e}^{t(\mathbf{A}+\alpha\mathbf{I})} = \mathrm{e}^{t\mathbf{A}}\mathrm{e}^{t\alpha\mathbf{I}} = \mathbf{B} \implies \mathbf{0} < \mathrm{e}^{-\alpha t}\mathbf{B} = \mathrm{e}^{t\mathbf{A}}.$$

Conversely, if $\mathbf{0} < \mathrm{e}^{t\mathbf{A}} = \sum_{k=0}^{\infty} t^k\mathbf{A}^k/k!$ for all $t > 0$, then $a_{ij} \geq 0$ for every $i \neq j$, for if $a_{ij} < 0$ for some $i \neq j$, then there exists a sufficiently small $t > 0$ such that $[\mathbf{I} + t\mathbf{A} + t^2\mathbf{A}^2/2 + \cdots]_{ij} < 0$, which is impossible. Furthermore, $\mathbf{A}$ must be irreducible; otherwise

$$\mathbf{A} \sim \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix} \implies \mathrm{e}^{t\mathbf{A}} = \sum_{k=0}^{\infty} t^k\mathbf{A}^k/k! \sim \begin{pmatrix} \star & \star \\ \mathbf{0} & \star \end{pmatrix}, \quad \text{which is impossible.}$$

8.3.13. (a) Being essentially positive implies that there exists some $\alpha \in \Re$ such that $\mathbf{A} + \alpha\mathbf{I}$ is nonnegative and irreducible (by Exercise 8.3.12). If $(r, \mathbf{x})$ is the Perron eigenpair for $\mathbf{A} + \alpha\mathbf{I}$, then for $\xi = r - \alpha$, $(\xi, \mathbf{x})$ is an eigenpair for $\mathbf{A}$.

(b) Every eigenvalue of $\mathbf{A} + \alpha\mathbf{I}$ has the form $z = \lambda + \alpha$, where $\lambda \in \sigma(\mathbf{A})$, so if r is the Perron root of $\mathbf{A} + \alpha\mathbf{I}$, then for $z \neq r$,

$$|z| < r \implies \operatorname{Re}(z) < r \implies \operatorname{Re}(\lambda + \alpha) < r \implies \operatorname{Re}(\lambda) < r - \alpha = \xi.$$

(c) If $\mathbf{A} \leq \mathbf{B}$, then $\mathbf{A} + \alpha\mathbf{I} \leq \mathbf{B} + \alpha\mathbf{I}$, so Wielandt's theorem (p. 675) insures that $r_A = \rho(\mathbf{A} + \alpha\mathbf{I}) \leq \rho(\mathbf{B} + \alpha\mathbf{I}) = r_B$, and hence $\xi_A = r_A - \alpha \leq r_B - \alpha = \xi_B$.

8.3.14. If $\mathbf{A}$ is primitive with $r = \rho(\mathbf{A})$, then, by (8.3.10) on p. 674,

$$\begin{aligned}
\left(\frac{\mathbf{A}}{r}\right)^k \to \mathbf{G} > \mathbf{0} &\implies \exists\, k_0 \text{ such that } \left(\frac{\mathbf{A}}{r}\right)^m > \mathbf{0} \quad \forall m \geq k_0 \\
&\implies \frac{a_{ij}^{(m)}}{r^m} > 0 \quad \forall m \geq k_0 \\
&\implies \lim_{m\to\infty} \left(\frac{a_{ij}^{(m)}}{r^m}\right)^{1/m} \to 1 \implies \lim_{m\to\infty} \left[a_{ij}^{(m)}\right]^{1/m} = r.
\end{aligned}$$

Conversely, we know from the Perron–Frobenius theorem that $r > 0$, so if $\lim_{k\to\infty}\left[a_{ij}^{(k)}\right]^{1/k} = r$, then $\exists\ k_0$ such that $\forall m \geq k_0$, $\left[a_{ij}^{(m)}\right]^{1/m} > 0$, which implies that $\mathbf{A}^m > \mathbf{0}$, and thus $\mathbf{A}$ is primitive by Frobenius's test (p. 678).

Solutions for exercises in section 8. 4

8.4.1. The left-hand Perron vector for $\mathbf{P}$ is $\boldsymbol{\pi}^T = (10/59,\ 4/59,\ 18/59,\ 27/59)$. It's the limiting distribution in the regular sense because $\mathbf{P}$ is primitive (it has a positive diagonal entry—recall Example 8.3.3 (p. 678)).

8.4.2. The left-hand Perron vector is $\boldsymbol{\pi}^T = (1/n)(1, 1, \ldots, 1)$. Thus the limiting distribution is the uniform distribution, and in the long run, each state is occupied an equal proportion of the time. The limiting matrix is $\mathbf{G} = (1/n)\mathbf{e}\mathbf{e}^T$.

8.4.3. If $\mathbf{P}$ is irreducible, then $\rho(\mathbf{P}) = 1$ is a simple eigenvalue for $\mathbf{P}$, so

$$rank\,(\mathbf{I}-\mathbf{P}) = n - \dim N\,(\mathbf{I}-\mathbf{P}) = n - geo\ mult_{\mathbf{P}}\,(1) = n - alg\ mult_{\mathbf{P}}\,(1) = n-1.$$

8.4.4. Let $\mathbf{A} = \mathbf{I}-\mathbf{P}$, and recall that $rank\,(\mathbf{A}) = n-1$ (Exercise 8.4.3). Consequently,

$$\mathbf{A} \text{ singular} \Longrightarrow \mathbf{A}[\operatorname{adj}(\mathbf{A})] = \mathbf{0} = [\operatorname{adj}(\mathbf{A})]\mathbf{A} \quad \text{(Exercise 6.2.8, p. 484)},$$

and

$$rank\,(\mathbf{A}) = n-1 \Longrightarrow rank\,(\operatorname{adj}(\mathbf{A})) = 1 \quad \text{(Exercises 6.2.11)}.$$

It follows from $\mathbf{A}[\operatorname{adj}(\mathbf{A})] = \mathbf{0}$ and the Perron–Frobenius theorem that each column of $[\operatorname{adj}(\mathbf{A})]$ must be a multiple of $\mathbf{e}$ (the column of 1's or, equivalently, the right-hand Perron vector for $\mathbf{P}$), so $[\operatorname{adj}(\mathbf{A})] = \mathbf{e}\mathbf{v}^T$ for some vector $\mathbf{v}$. But $[\operatorname{adj}(\mathbf{A})]_{ii} = P_i$ forces $\mathbf{v}^T = (P_1,\ P_2,\ \ldots,\ P_n)$. Similarly, $[\operatorname{adj}(\mathbf{A})]\mathbf{A} = \mathbf{0}$ insures that each row in $[\operatorname{adj}(\mathbf{A})]$ is a multiple of $\boldsymbol{\pi}^T$ (the left-hand Perron vector of $\mathbf{P}$), and hence $\mathbf{v}^T = \alpha\boldsymbol{\pi}^T$ for some α. This scalar α can't be zero; otherwise $[\operatorname{adj}(\mathbf{A})] = \mathbf{0}$, which is impossible because $rank\,(\operatorname{adj}(\mathbf{A})) = 1$. Therefore, $\mathbf{v}^T\mathbf{e} = \alpha \neq 0$, and $\mathbf{v}^T/(\mathbf{v}^T\mathbf{e}) = \mathbf{v}^T/\alpha = \boldsymbol{\pi}^T$.

8.4.5. If $\mathbf{Q}_{k\times k}$ $(1 \leq k < n)$ is a principal submatrix of $\mathbf{P}$, then there is a permutation matrix $\mathbf{H}$ such that $\mathbf{H}^T\mathbf{P}\mathbf{H} = \begin{pmatrix} \mathbf{Q} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{pmatrix} = \widetilde{\mathbf{P}}$. If $\mathbf{B} = \begin{pmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, then $\mathbf{B} \leq \widetilde{\mathbf{P}}$, and we know from Wielandt's theorem (p. 675) that $\rho(\mathbf{B}) \leq \rho\left(\widetilde{\mathbf{P}}\right) = 1$, and if $\rho(\mathbf{B}) = \rho\left(\widetilde{\mathbf{P}}\right) = 1$, then there is a number ϕ and a nonsingular diagonal matrix $\mathbf{D}$ such that $\mathbf{B} = \mathrm{e}^{\mathrm{i}\phi}\mathbf{D}\widetilde{\mathbf{P}}\mathbf{D}^{-1}$ or, equivalently, $\widetilde{\mathbf{P}} = \mathrm{e}^{-\mathrm{i}\phi}\mathbf{D}\mathbf{B}\mathbf{D}^{-1}$. But this implies that $\mathbf{X} = \mathbf{0}$, $\mathbf{Y} = \mathbf{0}$, and $\mathbf{Z} = \mathbf{0}$, which is impossible because $\mathbf{P}$ is irreducible. Therefore, $\rho(\mathbf{B}) < 1$, and thus $\rho(\mathbf{Q}) < 1$.

8.4.6. In order for $\mathbf{I}-\mathbf{Q}$ to be an M-matrix, it must be the case that $[\mathbf{I}-\mathbf{Q}]_{ij} \leq 0$ for $i \neq j$, and $\mathbf{I}-\mathbf{Q}$ must be nonsingular with $(\mathbf{I}-\mathbf{Q})^{-1} \geq \mathbf{0}$. It's clear that $[\mathbf{I}-\mathbf{Q}]_{ij} \leq 0$ because $0 \leq q_{ij} \leq 1$. Exercise 8.4.5 says that $\rho(\mathbf{Q}) < 1$, so

the Neumann series expansion (p. 618) insures that $\mathbf{I}-\mathbf{Q}$ is nonsingular and $(\mathbf{I}-\mathbf{Q})^{-1}=\sum_{j=1}^{\infty}\mathbf{Q}^j \geq \mathbf{0}$. Thus $\mathbf{I}-\mathbf{Q}$ is an M-matrix.

8.4.7. We know from Exercise 8.4.6 that every principal submatrix of order $1 \leq k < n$ is an M-matrix, and M-matrices have positive determinants by (7.10.28) on p. 626.

8.4.8. You can consider an absorbing chain with eight states

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

similar to what was described in Example 8.4.5, or you can use a four-state chain in which the states are defined to be the *number* of controls that hold at each activation of the system. Using the eight-state chain yields the following mean-time-to-failure vector.

$$\begin{array}{c} (1,1,1)\\ (1,1,0)\\ (1,0,1)\\ (0,1,1)\\ (1,0,0)\\ (0,1,0)\\ (0,0,1) \end{array} \begin{pmatrix} 368.4\\ 366.6\\ 366.6\\ 366.6\\ 361.3\\ 361.3\\ 361.3 \end{pmatrix} = (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e}.$$

8.4.9. This is a Markov chain with nine states (c,m) in which c is the chamber occupied by the cat, and m is the chamber occupied by the mouse. There are three absorbing states—namely $(1,1)$, $(2,2)$, $(3,3)$. The transition matrix is

$$\mathbf{P}=\frac{1}{72}\begin{array}{c} \\ (1,2)\\ (1,3)\\ (2,1)\\ (2,3)\\ (3,1)\\ (3,2)\\ (1,1)\\ (2,2)\\ (3,3) \end{array} \begin{array}{c} \begin{array}{ccccccccc} (1,2) & (1,3) & (2,1) & (2,3) & (3,1) & (3,2) & (1,1) & (2,2) & (3,3) \end{array}\\ \begin{pmatrix} 18 & 12 & 3 & 6 & 3 & 9 & 6 & 9 & 6\\ 12 & 18 & 3 & 9 & 3 & 6 & 6 & 6 & 9\\ 3 & 3 & 18 & 9 & 12 & 6 & 6 & 9 & 6\\ 4 & 6 & 6 & 18 & 4 & 8 & 2 & 12 & 12\\ 3 & 3 & 12 & 6 & 18 & 9 & 6 & 6 & 9\\ 6 & 4 & 4 & 8 & 6 & 18 & 2 & 12 & 12\\ 0 & 0 & 0 & 0 & 0 & 0 & 72 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 72 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 72 \end{pmatrix} \end{array}$$

The expected number of steps until absorption and absorption probabilities are

$$(\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{e}= \begin{array}{c} (1,2)\\ (1,3)\\ (2,1)\\ (2,3)\\ (3,1)\\ (3,2) \end{array} \begin{pmatrix} 3.24\\ 3.24\\ 3.24\\ 2.97\\ 3.24\\ 2.97 \end{pmatrix} \quad \text{and} \quad (\mathbf{I}-\mathbf{T}_{11})^{-1}\mathbf{T}_{12}= \begin{array}{c} \begin{array}{ccc} (1,1) & (2,2) & (3,3) \end{array}\\ \begin{pmatrix} 0.226 & 0.41 & 0.364\\ 0.226 & 0.364 & 0.41\\ 0.226 & 0.41 & 0.364\\ 0.142 & 0.429 & 0.429\\ 0.226 & 0.364 & 0.41\\ 0.142 & 0.429 & 0.429 \end{pmatrix} \end{array}$$

Index

A

absolute uncertainty or error, 414
absorbing Markov chains, 700
absorbing states, 700
addition, properties of, 82
additive identity, 82
additive inverse, 82
adjacency matrix, 100
adjoint, 84, 479
adjugate, 479
affine functions, 89
affine space, 436
algebraic group, 402
algebraic multiplicity, 496, 510
amplitude, 362
Anderson-Duffin formula, 441
Anderson, Jean, xii
Anderson, W. N., Jr., 441
angle, 295
  canonical, 455
  between complementary spaces, 389, 450
  maximal, 455
  principal, 456
  between subspaces, 450
annihilating polynomials, 642
aperiodic Markov chain, 694
Arnoldi's algorithm, 653
Arnoldi, Walter Edwin, 653
associative property
  of addition, 82
  of matrix multiplication, 105
  of scalar multiplication, 83
asymptotic rate of convergence, 621
augmented matrix, 7
Autonne, L., 411

B

back substitution, 6, 9
backward error analysis, 26
backward triangle inequality, 273
band matrix, 157
Bartlett, M. S., 124
base-b, 375
basic columns, 45, 61, 178, 218
  combinations of, 54
basic variables, 58, 61
  in nonhomogeneous systems, 70
basis, 194, 196
  change of, 253
  characterizations, 195
  orthonormal, 355
basis for
  direct sum, 383
  intersection of spaces, 211
  space of linear transformations, 241
Bauer–Fike bound, 528
beads on a string, 559
Bellman, Richard, xii
Beltrami, Eugenio, 411
Benzi, Michele, xii
Bernoulli, Daniel, 299
Bessel, Friedrich W., 305
Bessel's inequality, 305
best approximate inverse, 428
biased estimator, 446
binary representation, 372
Birkhoff, Garrett, 625
Birkhoff's theorem, 702
bit reversal, 372
bit reversing permutation matrix, 381
block diagonal, 261–263
  rank of, 137
  using real arithmetic, 524
block matrices, 111
  determinant of, 475
  linear operators, 392
block matrix multiplication, 111
block triangular, 112, 261–263
  determinants, 467
  eigenvalues of, 501
Bolzano–Weierstrass theorem, 670
Boolean matrix, 679
bordered matrix, 485
  eigenvalues of, 552
branch, 73, 204
Brauer, Alfred, 497
Bunyakovskii, Victor, 271

C

cancellation law, 97
canonical angles, 455
canonical form, reducible matrix, 695
Cantor, Georg, 597
Cauchy, Augustin-Louis, 271
  determinant formula, 484
  integral formula, 611
Cauchy–Bunyakovskii–Schwarz inequality, 272
Cauchy–Goursat theorem, 615
Cauchy–Schwarz inequality, 287
Cayley, Arthur, 80, 93, 158, 460
Cayley–Hamilton theorem, 509, 532, 597
  to determine $f(\mathbf{A})$, 614
Cayley transformation, 336, 556
CBS inequality, 272, 277, 473
  general form, 287
centered difference approximations, 19
Cesàro, Ernesto, 630
Cesàro sequence, 630
Cesàro summability, 630, 633, 677
  for stochastic matrix, 697
chain, Jordan, 575
change of basis, 251, 253, 258

change of coordinates, 252
characteristic equation, 491, 492
  coefficients, 494, 504
characteristic polynomial, 491, 492
  of a product, 503
characteristic values and vectors, 490
Chebyshev, Pafnuty Lvovich, 40, 687
checking an answer, 35, 416
Cholesky, Andre-Louis, 154
Cholesky factorization, 154, 314, 558, 559
Cimmino, Gianfranco, 445
Cimmino's reflection method, 445
circuit, 204
circulant matrix, 379
  with convolution, 380
  eigenvalues, eigenvectors, 523
classical Gram–Schmidt algorithm, 309
classical least squares, 226
clock cycles, 539, 694
closest point
  to an affine space, 436
  with Fourier expansion, 440
  theorem, 435
closure property
  of addition, 82, 160
  of multiplication, 83, 160
coefficient matrix, 7, 42
coefficient of linear correlation, 297
cofactor, 477, 487
  expansion, 478, 481
Collatz, Lothar, 666
Collatz–Wielandt formula, 666, 673, 686
column, 7
  equivalence, 134
    and nullspace, 177
  operations, 14, 134
  rank, 198
  relationships, 50, 136
  scaling, 27
  space, 170, 171, 178
    spanning set for, 172
  vector, 8, 81
Comdico, David, xii
commutative law, 97
commutative property of addition, 82
commuting matrices, eigenvectors, 503, 522
companion matrix, 648
compatibility of norms, 285
compatible norms, 279, 280
competing species model, 546
complementary projector, 386
complementary subspaces, 383, 403
  angle between, 389, 450
complete pivoting, 28
  numerical stability, 349
complete set of eigenvectors, 507
complex conjugate, 83
complex exponential, 362, 544
complex numbers, the set of, 81
component matrices, 604
component vectors, 384
composition
  of linear functions, 93
  of linear transformations, 245, 246
  of matrix functions, 608, 615
computer graphics, 328, 330
condition number
  for eigenvalues, 528
  generalized, 426
  for matrices, 127, 128, 414, 415
condition of
  eigenvalues, hermitian matrices, 552
  linear system, 128
conditioning and pivots, 426
conformable, 96
conformably partitioned, 111
congruence transformation, 568
conjugate, complex, 83
conjugate gradient algorithm, 657
conjugate matrix, 84
conjugate transpose, 84
  reverse order law, 109
conjugate vectors, 657
connected graph, 202
connectivity and linear dependence, 208
connectivity matrix, 100
consistent system, 53, 54
constituent matrices, 604
continuity
  of eigenvalues, 497
  of inversion, 480
  of norms, 277
continuous Fourier transform, 357
continuous functions, max and min, 276
convergence, 276, 277
convergent matrix, 631
converse of a statement, 54
convolution
  with circulants, 380
  definition, 366
  operation count, 377
  theorem, 367, 368, 377
Cooley, J. W., 368, 375, 651
cooperating species model, 546
coordinate matrix, 242
coordinates, 207, 240, 299
  change of, 252
  of a vector, 240
coordinate spaces, 161
core-nilpotent decomposition, 397
correlation, 296
correlation coefficient, 297
cosine
  of angle, 295
    minimal angle, 450
  discrete, 361

Courant–Fischer theorem, 550
alternate, 557
for singular values, 555
Courant, Richard, 550
covariance, 447
Cramer, Gabriel, 476
Cramer's rule, 459, 476
critical point, 570
cross product, 332, 339
Cuomo, Kelly, xii
curve fitting, 186, 229

D

defective, 507
deficient, 496, 507
definite matrices, 559
deflation, eigenvalue problems, 516
dense matrix, 350
dependent set, 181
derivative
of a determinant, 471, 474, 486
of a linear system, 130
of a matrix, 103, 226
operator, 245
determinant, 461
computing, 470
of a product, 467
as product of eigenvalues, 494
of a sum, 485
and volume, 468
deviation from symmetry, 436
diagonal dominance, 639
diagonal matrix, 85
eigenvalues of, 501
inverse of, 122
diagonalizability, 507
being arbitrarily close to, 533
in terms of minimum polynomial, 645
in terms of multiplicities, 512
summary, 520
diagonalization
of circulants, 379
Jacobi's method, 353
of normal matrices, 547
simultaneous, 522
diagonally dominant, 184, 499, 622, 623, 639
systems, 193
difference equations, 515, 616
difference of matrices, 82
difference of projectors, 393
differential equations, 489, 541, 542
independent solutions, 481
nonhomogeneous, 609
solution of, 546
stability, 544, 609
systems, 608
uncoupling, 559
diffusion equation, 563
diffusion model, 542
dimension, 196
of direct sum, 383
of fundamental subspaces, 199
of left-hand nullspace, 218
of nullspace, 218
of orthogonal complement, 339
of range, 218
of row space, 218
of space of linear transformations, 241
of subspace, 198
of sum, 205
direct product, 380, 597
direct sum, 383
of linear operators, 399
of several subspaces, 392
of symmetric and skew-symmetric matrices, 391
directed distance between subspaces, 453
directed graph, 202
Dirichlet, Johann P. G. L, 563, 597
Dirichlet problem, 563
discrete Fourier transform, 356, 358
discrete Laplacian, 563
eigenvalues of, 598
discrete sine, cosine, and exponential, 361
disjoint subspaces, 383
distance, 271
to lower-rank matrices, 417
between subspaces, 450
to symmetric matrices, 436
between a vector and a subspace, 435
distinct eigenvalues, 514
distributions, 532
distributive property
of matrix multiplication, 105
of scalar multiplication, 83
domain, 89
doubly stochastic, 702
Drazin generalized inverse, 399, 401, 422, 640
Cauchy formula, 615
integral representation, 441
Drazin, M. P., 399
Duffin, R. J., 441
Duncan, W. J., 124

E

Eckart, C., 411
economic input–output model, 681
edge matrix, 331
edges, 202
eigenpair, 490
eigenspace, 490

eigenvalues, 266, 410, 490
bounds for, 498
continuity of, 497
determinant and trace, 494
distinct, 514
generalized, 571
index of, 401, 587, 596
perturbations and condition of, 528, 551
semisimple, 596
sensitivity, hermitian matrices, 552
unit, 696
eigenvalues of
bordered matrices, 552
discrete Laplacian, 566, 598
triangular and diagonal matrices, 501
tridiagonal Toeplitz matrices, 514
eigenvectors, 266, 490
of commuting matrices, 503
generalized, 593, 594
independent, 511
of tridiagonal Toeplitz matrices, 514
electrical circuits, 73, 204
elementary matrices, 131–133
interchange matrices, 135, 140
elementary orthogonal projector, 322, 431
rank of, 337
elementary reflector, 324, 444
determinant of, 485
elementary row and column operations, 4, 8
and determinants, 463
elementary triangular matrix, 142
ellipsoid, 414
degenerate, 425
elliptical inner product, 286
elliptical norm, 288
EP matrices, 408
equal matrices, 81
equivalence, row and column, 134
testing for, 137
equivalent norms
matrices, 425
vectors, 276
equivalent statements, 54
equivalent systems, 3
ergodic class, 695
error, absolute and relative, 414
essentially positive matrix, 686
estimators, 446
euclidean norm, 270
unitary invariance of, 321
evolutionary processes, 616
exponential
complex, 544
discrete, 361
matrix, 441, 525
inverse of, 614
products of, 539
sums of, 614
extending to a basis, 201
extending to an orthonormal basis, 325, 335, 338, 404
extension set, 188

F

Faddeev and Sominskii, 504
fail-safe system, 701
fast Fourier transform (FFT), 368
FFT algorithm, 368, 370, 373, 381, 651
FFT operation count, 377
fast integer multiplication, 375
filtering random noise, 418
finite difference matrix, 522, 639
finite-dimensional spaces, 195
finite group, 676
Fischer, Ernst, 550
five-point difference equations, 564
fixed points, 386, 391
of a reflector, 338
flatness, 164
floating-point number, 21
forward substitution, 145
four fundamental subspaces, 169
summary, 178
Fourier coefficients, 299
Fourier expansion, 299
and projection, 440
Fourier, Jean Baptiste Joseph, 299
Fourier matrix, 357
Fourier series, 299, 300
Fourier transform
continuous, 357
discrete, 356, 358
Frame, J. S., 504
Francis, J. F. G., 535
Fréchet, Maurice, R., 289
free variables, 58, 61
in nonhomogeneous systems, 70
frequency, 362
frequency domain, 363
Frobenius, Ferdinand Georg, 44, 123, 215, 662
Frobenius form, 680
Frobenius inequality, 221
Frobenius matrix norm, 279, 425, 428
and inner product, 288
of rank-one matrices, 391
unitary invariance of, 337
Frobenius test for primitivity, 678
full-rank factorization, 140, 221, 633
for determining index, 640
of a projector, 393
function
affine, 89
composition of, 93, 615, 608
domain of, 89
linear, 89, 238
norm of, 288
range of, 89

functional matrix identities, 608
functions of
  diagonalizable matrices, 526
  of Jordan blocks, 600
  matrices, 601
    using Cauchy integral formula, 611
    using Cayley–Hamilton theorem, 614
  nondiagonalizable matrices, 603
fundamental (normal) mode of vibration, 562
fundamental problem of matrix theory, 506
fundamental subspaces, 169
  dimension of, 199
  orthonormal bases for, 407
  projector onto, 434
fundamental theorem of algebra, 185, 492
fundamental theorem of linear algebra, 405

G

gap, 453, 454
Gauss, Carl F., ix, 2, 93, 234, 488
  as a teacher, 353
Gaussian elimination, 2, 3
  and LU factorization, 141
  effects of roundoff, 129
  modified, 43
  numerical stability, 348
  operation counts, 10
Gaussian transformation, 341
Gauss–Jordan method, 15, 47, 48
  for computing a matrix inverse, 118
  operation counts, 16
Gauss–Markov theorem, 229, 448
Gauss–Seidel method, 622
general solution
  algebraic equations
    homogeneous systems, 59, 61,
    nonhomogeneous systems, 64, 66, 70, 180, 221
  difference equations, 616
  differential equations, 541, 609
generalized condition number, 426
generalized eigenvalue problem, 571
generalized eigenvectors, 593, 594
generalized inverse, 221, 393, 422, 615
  Drazin, 399
  group, 402
  and orthogonal projectors, 434
generalized minimal residual (GMRES), 655
genes and chromosomes, 543
geometric multiplicity, 510
geometric series, 126, 527, 618
Gerschgorin circles, 498
Gerschgorin, S. A., 497
Givens reduction, 344
  and determinants, 485
  numerical stability, 349
Givens rotations, 333
Givens, Wallace, 333
GMRES, 655
Golub, Gene H., xii
gradient, 570
Gram, Jorgen P., 307
Gram matrix, 307
Gram–Schmidt algorithm
  classical version, 309
  implementations of, 319
  and minimum polynomial, 643
  modified version, 316
  numerical stability of, 349
  and volume, 431
Gram–Schmidt process, 345
Gram–Schmidt sequence, 308, 309
graph, 202
  of a matrix, 209, 671
graphics, 3-D rotations, 328, 330
Grassmann, Hermann G., 160
Graybill, Franklin A., xii
grid norm, 274
grid points, 18
group, finite, 676
group inverse, 402, 640, 641
growth in Gaussian elimination, 26
Guttman, L., 124

H

Hadamard, Jacques, 469, 497
Hadamard's inequality, 469
Halmos, Paul, xii, 268
Hamilton, William R., 509
harmonic functions, 563
Haynsworth, Emilie V., 123
heat equation, 563
Helfrich, Laura, xii
Hermite, Charles, 48
Hermite interpolation polynomial, 607
Hermite normal form, 48
Hermite polynomial, 231
hermitian matrix, 85, 409, 410
  condition of eigenvalues, 552
  eigen components of, 549
Hessenberg matrices 350
  QR factorization of, 352
Hessian matrix, 570
Hestenes, Magnus R., 656
hidden surfaces, 332, 339
Hilbert, David, 307
Hilbert matrix, 14, 31, 39
Hilbert–Schmidt norm, 279
Hohn, Franz, xii
Hölder, Ludwig O., 278
Hölder's inequality, 274, 277, 278
homogeneous systems, 57, 61
Hooke, Robert, 86
Hooke's law, 86
Horn, Roger, xii

Horst, Paul, 504
Householder, Alston S., 324
Householder reduction, 341, 342
  and determinants, 485
  and fundamental subspaces, 407
  numerical stability, 349
Householder transformations, 324
hyperplane, 442

I

idempotent, 113, 258, 339, 386
  and projectors, 387
identity matrix, 106
identity operator, 238
ill-conditioned matrix, 127, 128, 415
ill-conditioned system, 33, 535
  normal equations, 214
image and image space, 168, 170
  dimension of, 208
image of unit sphere, 417
imaginary, pure, 556
imprimitive matrices, 674
  maximal root of, 676
  spectrum of, 677
  test for, 678
imprimitivity, index of, 679, 680
incidence matrix, 202
inconsistent system, 53
independent columns, 218
independent eigenvectors, 511
independent rows, 218
independent set, 181
  basic facts, 188
  maximal, 186
independent solutions
  for algebraic equations, 209
  for differential equations, 481
index
  of an eigenvalue, 401, 587, 596
  of imprimitivity, 674, 679, 680
  of nilpotency, 396
  of a square matrix, 394, 395
    by full-rank factorization, 640
induced matrix norm, 280, 389
  of $\mathbf{A}^{-1}$, 285
  elementary properties, 285
  of rank-one matrices, 391
  unitary invariance of, 337
inertia, 568
infinite-dimensional spaces, 195
infinite series and matrix functions, 527
infinite series of matrices, 605
information retrieval, 419
inner product, 286
  geometric interpretation, 431
input–output economic model, 681
integer matrices, 156, 473, 485
integer multiplication, 375
integral formula
  for generalized inverses, 441
  for matrix functions, 611
intercept model, 447
interchange matrices, 135, 140
interlacing of eigenvalues, 552
interpolation
  formula for $f(\mathbf{A})$, 529
  Hermite polynomial, 607
  Lagrange polynomial, 186
intersection of subspaces
  basis for, 211
  projection onto, 441
invariant subspace, 259, 262, 263
inverse Fourier transform, 358
inverse iteration, 534
inverse matrix, 115
  best approximation to, 428
  Cauchy formula for, 615
  computation of, 118
    operation counts, 119
  continuity of, 480
  determinants, 479
  eigenvalues of, 501
  existence of, 116
  generalized, 615
  integral representation of, 441
  norm of, 285
  properties of, 120
  of a sum, 220
invertible operators, 246, 250
invertible part of an operator, 399
involutory, 113, 325, 339, 485
irreducible Markov chain, limits, 693
irreducible matrix, 209, 671
isometry, 321
iteration matrix, 620
iterative methods, 620

J

Jacobi's diagonalization method, 353
Jacobi's iterative method, 622, 626
Jacobi, Karl G. J., 353
Johnson, Charlie, xii
Jordan blocks, 588, 590
  functions of, 600
  nilpotent, 579
Jordan chains, 210, 401, 575, 576, 593
  construction of, 594
Jordan form, 397, 408, 589, 590
  for nilpotent matrices, 579
  preliminary version, 397
Jordan, Marie Ennemond Camille, 15, 411, 589
Jordan segment, 588, 590
Jordan structure of matrices, 580, 581, 586, 589
  uniqueness of, 580

Jordan, Wilhelm, 15

K

Kaczmarz's projection method, 442, 443
Kaczmarz, Stefan, 442
Kaplansky, Irving, 268
Kearn, Vickie, xi, 12
kernel, 173
Kirchhoff's rules, 73
  loop rule, 204
Kline, Morris, 80
Kowa, Seki, 459
Kronecker, Leopold, 597
Kronecker product, 380, 597
  and the Laplacian, 573
Krylov, Aleksei Nikolaevich, 645
Krylov
  method, 649
  sequence, 401
  subspaces, sequences, matrices, 646
Kummer, Ernst Eduard, 597

L

Lagrange interpolating polynomial, 186, 230, 233, 529
Lagrange, Joseph-Louis, 186, 572
Lagrange multipliers, 282
Lancaster, Peter, xii
Lanczos algorithm, 651
Lanczos, Cornelius, 651
Laplace's determinant expansion, 487
Laplace's equation, 624
Laplace, Pierre-Simon, 81, 307, 487, 572
Laplacian, 563
latent semantic indexing, 419
latent values and vectors, 490
law of cosines, 295
LDU factorization, 154
leading principal minor, 558
leading principal submatrices, 148, 156
least common multiple, 647
least squares, 226, 439
  and Gram–Schmidt, 313
  and orthogonal projection, 437
  and polynomial fitting, 230
  and pseudoinverse, 438
  and QR factorization, 346
  total least squares, 223
  why least squares?, 446
LeBlanc, Kathleen, xii
left-hand eigenvectors, 490, 503, 516, 523, 524
  in inverses, 521
  and projectors, 518
left-hand nullspace, 174, 178, 199
  spanning set for, 176
Legendre, Adrien–Marie, 319, 572
Legendre polynomials, 319
Legendre's differential equation, 319
Leibniz, Gottfried W., 459
length of a projection, 323
Leontief's input–output model, 681
Leontief, Wassily, 681
Leslie, P. H., 684
Leslie population model, 683
Leverrier–Souriau–Frame Algorithm, 504
Leverrier, U. J. J., 504
Lévy, L., 497
limiting distribution, 531, 636
limits
  and group inversion, 640
  in Markov chains
    irreducible Markov chains, 693
    reducible Markov chains, 698
  of powers of matrices, 630
  and spectral radius, 617
  of vector sequences, 639
  in vector spaces, 276, 277
Lindemann, Carl Louis Ferdinand von, 662
linear
  algebra, 238
  combination, 91
  correlation, 296, 306
  dependence and connectivity, 208
  estimation, 446
  functions, 89, 238
    defined by matrix multiplication, 106
    defined by systems of equations, 99
  models, 448
  operators, 238
    and block matrices, 392
  regression, 227, 446
  spaces, 169
  stationary iterations, 620
  transformation, 238
linearly independent and dependent sets, 181
  basic facts, 188
  maximal, 186
  and rank, 183
linearly independent eigenvectors, 511
lines in $\Re^n$ not through the origin, 440
lines, projection onto, 440
long-run distribution, 531
loop, 73
  equations, 204
  rule, 74
  simple, 75
lower triangular, 103
LU factorization, 141, 144
  existence of, 149
  with interchanges, 148
  operation counts, 146
  summary, 153

M

main diagonal, 41, 85
Markov, Andrei Andreyevich, 687
Markov chains, 532, 638, 687
  absorbing, 700
  periodic, 694
mass-stiffness equation, 571
matrices, the set of, 81
matrix, 7
  diagonal 85
  exponential, 441, 525, 529
    and differential equations, 541, 546, 608
    inverse of, 614
    products, 539
    sums, 614
  functions, 526, 601
    as infinite series, 527
    as polynomials, 606
  group, 402
  multiplication, 96
    by blocks, 111
    as a linear function, 106
    properties of, 105
    relation to linear transformations, 244
  norms, 280
    1-norm, 283
    2-norm, 281, 425
    ∞-norm, 127, 283
    Frobenius norm, 425
    induced norm, 285
  polynomials, 501
  product, 96
  representation of a projector, 387
  representations, 262
  triangular 41
maximal angle, 455
maximal independent set, 218
maximal linearly independent subset, 186, 196
maximum and minimum of continuous functions, 276
McCarthy, Joseph R., 651
mean, 296, 447
Meyer
  Bethany B., xii
  Carl, Sr., xii
  Holly F., xii
  Louise, xii
  Margaret E., xii
  Martin D., xii
min-max theorem, 550
  alternate formulation, 557
  for singular values, 555
minimal angle, 450
minimal spanning set, 196, 197
minimum norm least squares solution, 438
minimum norm solution, 426
minimum polynomial, 642
  determination of, 643
  of a vector, 646
minimum variance estimator, 446
Minkowski, Hermann, 184, 278, 497, 626
Minkowski inequality, 278
minor determinant, principal, 559, 466
MINRES algorithm, 656
Mirsky, Leonid, xii
M-matrix, 626, 639, 682, 703
modern least squares, 437
modified gaussian elimination, 43
modified Gram–Schmidt algorithm, 316
monic polynomial, 642
Montgomery, Michelle, xii
Moore, E. H., 221
Moore–Penrose generalized inverse, 221, 422, 400
  best approximate inverse, 428
  integral representation, 441
  and orthogonal projectors, 434
Morrison, W. J., 124
multiplication
  of integers, 375
  of matrices, 96
  of polynomials, 367
multiplicities, 510
  and diagonalizability, 512
multiplier, 22, 25
  in partial pivoting, 26

N

negative definite, 570
Neumann series, 126, 527, 618
Newton, 86
Newton's identities, 504
Newton's second law, 560
nilpotent, 258, 396, 502, 510
  Jordan blocks, 579
  part of an operator, 399
Noble, Ben, xii
node, 18, 73, 202, 204
  rule, 74, 204
no-intercept model, 447
noise removal with SVD, 418
nonbasic columns, 50, 61
nonderogatory matrices, 644, 648
nondiagonalizable, spectral resolution, 603
nonhomogeneous differential equations, 609
nonhomogeneous systems, 57, 64
  general solution, 64, 66, 70
  summary, 70
nonnegative matrices, 661, 670
nonsingular matrices, 115
  and determinants, 465
  and elementary matrices, 133
  products of, 121
  sequences of, 220

norm, 269
compatibility, 279, 280, 285
elliptical, 288
equivalent, 276, 425
of a function, 288
on a grid, 274
of an inverse, 285
for matrices, 280
1-, 2-, and ∞-norms, 281, 283
Frobenius, 279, 337
induced, 280, 285, 337
of a projection, 323
for vectors, 275
1-, 2-, and ∞-norms, 274
p-norms, 274
of a waveform, 382
normal equations, 213, 214, 221, 226, 313, 437
normal modes of vibration, 562, 571
normalized vector, 270
normal matrix, 304, 400, 409, 547
nullity, 200, 220
nullspace, 173, 174, 178, 199
equality, 177
of an orthogonal projector, 434
of a partitioned matrix, 208
of a product, 180, 220
spanning set for, 175
and transpose, 177
number of pivots, 218
numerical stability, 347

O

oblique projection, 385
method for linear systems, 443
oblique projectors from SVD, 634
Ohm's law, 73
Oh notation $O(h^p)$, 18
one-to-one mapping, 250
onto mapping, 250
operation counts
for convolution, 377
for Gaussian elimination, 10
for Gauss–Jordan method, 16
for LU factorization, 146
for matrix inversion, 119
operator, linear, 238
operator norm, 280
Ortega, James, xii
orthogonal complement, 322, 403
dimension of, 339
involving range and nullspace, 405
orthogonal decomposition theorem, 405, 407
orthogonal diagonalization, 549
orthogonal distance, 435
orthogonal matrix, 320
determinant of, 473
orthogonal projection, 239, 243, 248, 299, 305, 385, 429
and 3-D graphics, 330
onto an affine space, 436
and least squares, 437
orthogonal projectors, 322, 410, 427, 429
elementary, 431
formulas for, 430
onto an intersection, 441
and pseudoinverses, 434
sums of, 441
orthogonal reduction, 341
to determine full-rank factorization, 633
to determine fundamental subspaces, 407
orthogonal triangularization, 342
orthogonal vectors, 294
orthonormal basis, 298
extending to, 325, 335, 38
for fundamental subspaces, 407
by means of orthogonal reduction, 355
orthonormal set, 298
Ostrowski, Alexander, 626
outer product, 103
overrelaxation, 624

P

Painter, Richard J., xii
parallelepiped, 431, 468
parallelogram identity, 290, 291
parallelogram law, 162
parallel sum, 441
parity of a permutation, 460
Parseval des Chênes, M., 305
Parseval's identity, 305
partial pivoting, 24
and diagonal dominance, 193
and LU factorization, 148
and numerical stability, 349
particular solution, 58, 65–68, 70, 180, 213
partitioned matrix, 111
and linear operators, 392
rank and nullity of, 208
Peano, Giuseppe, 160
Penrose equations, 422
Penrose, Roger, 221
perfect shuffle, 372, 381
period of trigonometric functions, 362
periodic extension, 302
periodic function, 301
periodic Markov chain, 694
permutation, 460
symmetric, 671
permutation counter, 151
permutation matrix, 135, 140, 151
perpendicular, 294
perp, properties of, 404, 409
Perron–Frobenius theory, 661, 673
Perron, Oskar, 661
Perron root, 666, 668

Perron vector, 665, 668, 673
perturbations
affecting rank, 216
eigenvalues, 528
hermitian eigenvalues, 551
in inverses, 128
in linear systems, 33, 128, 217
rank-one update, 208
singular values, 421
Piazzi, Giuseppe, 233
pivot
conditioning, 426
determinant formula for, 474, 558
elements and equations, 5
positions, 5, 58, 61
in partial pivoting, 24
uniqueness, 44
pivoting
complete, 28
partial, 24
plane rotation, 333
determinant of, 485
p-norm, 274
Poisson's equation, 563, 572
Poisson, Siméon D., 78, 572
polar factorization, 572
polarization identity, 293
polynomial
equations, 493
in a matrix, 501
and matrix functions, 606
minimum, 642
multiplication and convolution, 367
polytope, 330, 339
ponderal index, 236
poor man's root finder, 649
population distribution, 532
population migration, 531
population model, Leslie, 683
positive definite form, 567
positive definite matrix, 154, 474, 558, 559
positive matrix, 661, 663
positive semidefinite matrix, 558, 566
Poulson, Deborah , xii
power method, 532, 533
powers of a matrix, 107
limiting values, 530
powers of linear transformations, 248
precision, 21
preconditioned system, 658
predator–prey model, 544
primitive matrices, 674
test for, 678
principal angles, 456
principal minors, 494, 558
in an M-matrix, 626, 639
nonnegative, 566
positive, 559
principal submatrix, 494, 558
and interlaced eigenvalues, 553
of an M-matrix, 626
of a stochastic, 703
products
of matrices, 96
of nonsingular matrices, 121
of orthogonal projectors, 441
of projectors, 393
product rule for determinants, 467
projection, 92, 94, 322, 385, 429
and Fourier expansion, 440
method for solving linear systems, 442, 443
onto
affine spaces, 436
fundamental subspaces, 434
hyperplanes, 442
lines, 440, 431
oblique subspaces, 385
orthogonal subspaces, 429
symmetric matrices, 436
projectors, 239, 243, 339, 385, 386
complementary, 386
from core-nilpotent decomposition, 398
difference of, 393
from full-rank factorization, 633, 634
as idempotents, 387
induced norm of, 389
matrix representation of, 387
oblique, 386
orthogonal, 429
product of, 393
spectral, 517, 603
sum of, 393
proper values and vectors, 490
pseudoinverse, 221, 422, 615
as best approximate inverse, 428
Drazin, 399
group, 402
inner, outer, reflexive, 393
integral representation of, 441, 615
and least squares, 438
Moore–Penrose, 422
and orthogonal projectors, 434
pure imaginary, 556
Pythagorean theorem, 294, 305, 423
and closest point theorem, 435
for matrices with Frobenius norm, 428

Q

QR factorization, 345, 535
and Hessenberg matrices, 352
and least squares, 346
and minimum polynomial, 643
rectangular version of, 311
and volume, 431
quadratic form, 567

quaternions, 509

R

random integer matrices, 156
random walk, 638
range
  of a function, 89, 169
  of a matrix, 170, 171, 178, 199
  of an operator, 250
  of an orthogonal projector, 434
  of a partitioned matrix, 179
  of a product, 180, 220
  of a projector, 391
  of a sum, 206
range-nullspace decomposition, 394, 407
range-symmetric matrices, 408
rank, 45, 139
  of a block diagonal matrix, 137
  and consistency, 54
  and determinants, 466
  of a difference, 208
  of an elementary projector, 337
  and free variables, 61
  of an incidence matrix, 203
  and independent sets, 183
  and matrix inverses, 116
  and nonhomogeneous systems, 70
  and nonsingular submatrices, 218
  numerical determination, 421
  of a partitioned matrix, 208
  of a perturbed matrix, 216
  of a product, 210, 211, 219
  of a projector, 392
  and submatrices, 215
  of a sum, 206, 221
  summary, 218
  and trivial nullspaces, 175
rank normal form, 136
rank-one matrices
  characterization of, 140
  diagonalizability of, 522
  perturbations of, 208
rank-one updates
  determinants of, 475
  eigenvalues of, 503
rank plus nullity theorem, 199, 410
Rayleigh, Lord, 550
Rayleigh quotient, 550
  iteration, 535
real numbers, the set of, 81
real Schur form, 524
real-symmetric matrix, 409, 410
rectangular matrix, 8
rectangular QR factorization, 311
rectangular systems, 41
reduced row echelon form, 48
reducible Markov chain, 698
reducible matrices, 209, 671
  canonical form for, 695
  in linear systems, 112
reflection, 92, 94
  about a hyperplane, 445
  method for solving linear systems, 445
reflector, 239, 324, 444
  determinant of, 485
reflexive pseudoinverse, 393
regression, 227, 446
relative uncertainty or error, 414
relaxation parameter, 445, 624
residual, 36, 416
resolvent, 285, 611
restricted operators, 259, 393, 399
restricted transformations, 424
reversal matrix, 596
reverse order law
  for inversion, 120, 121
  for transpose and conjugate transpose, 109
reversing binary bits, 372
Richardson iterative method, 622
right angle, 294
right-hand rule, 340
right-hand side, 3
Ritz values, 651
roots of unity, 356
  and imprimitive matrices, 676
Rose, Nick, xii
rotation, 92, 94
  determinant of, 485
  plane (Givens rotations), 333
  in $\Re^2$, 326
  in $\Re^3$, 328
  in $\Re^n$, 334
rotator, 239, 326
rounding convention, 21
roundoff error, 21, 129, 347
row, 7
  echelon form, 44
    reduced, 48
  equivalence, 134, 218
    and nullspace, 177
  operations, 134
  rank, 198
  relationships, 136
  scaling, 27
  space, 170, 171, 178, 199
    spanning set for, 172
  vector, 8, 81
RPN matrices, 408
Rutishauser, Heinz, 535

S

Saad, Yousef, 655
saw-toothed function, 306
scalar, 7, 81

scalar multiplication, 82, 83
scale, 27
scaling a linear system, 27, 28
scaling in 3-D graphics, 332
Schmidt, Erhard, 307
Schrödinger, Erwin, 651
Schultz, Martin H., 655
Schur complements, 123, 475
Schur form for real matrices, 524
Schur, Issai, 123, 508, 662
Schur norm, 279
Schur triangularization theorem, 508
Schwarz, Hermann A., 271, 307
search engine, 418, 419
sectionally continuous, 301
secular equation, 503
Seidel, Ludwig, 622
Sellers, Lois, xii
semiaxes, 414
semidefinite, 566
semisimple eigenvalue, 510, 591, 593, 596
semistable, 544
sensitivity, 128
  minimum norm solution, 426
sequence
  limit of, 639
  of matrices, 220
series for $f(\mathbf{A})$, 605
shape, 8
shell game, 635
Sherman, J., 124
Sherman–Morrison formula, 124, 130
SIAM, 324, 333
signal processing, 359
signal-to-noise ratio, 418
sign of a permutation, 461
similar matrices, 255, 473, 506
similarity, 505
  and block-diagonal matrices, 263
  and block-triangular matrices, 263
  and eigenvalues, 508
  invariant, 256
  and orthogonal matrices, 549
  transformation, 255, 408, 506
  and transpose, 596
  unitary, 547
simple eigenvalue, 510
simple loops, 75
simultaneous diagonalization, triangularization, 522
simultaneous displacements, 622
sine, discrete, 361
singular matrix, 115
  eigenvalues of, 501
  sequences of, 220
singular systems, practical solution of, 217
singular values, 553
  Courant–Fischer theorem, 555
  and determinants, 473
  as eigenvalues, 555
  and the SVD, 412
size, 8
skew-hermitian matrices, 85, 88
skew-symmetric matrices, 85, 88, 391, 473
  eigenvalues of, 549, 556
  as exponentials, 539
  vector space of, 436
SOR method, 624
Souriau, J. M., 504
spanning sets, 165
  for column space, 172
  for four fundamental subspaces, 178
  for left-hand nullspace, 176
  minimal, 197
  for nullspace, 175
  for row space, 172
  test for, 172
sparse least squares, 237
sparse matrix, 350
spectral circle, imprimitive matrices, 676
spectral mapping property, 539, 613
spectral projectors, 517, 602, 603
  commuting property, 522
  interpolation formula for, 529
  positivity of, 677
  in terms of eigenvectors, 518
spectral radius, 497, 521, 540
  Collatz–Wielandt formula, 666, 673, 686
  as a limit, 619
  and limits, 617
spectral representation of matrix functions, 526
spectral resolution of $f(\mathbf{A})$, 603
spectral theorem for diagonalizable matrices, 517
spectrum, 490
  of imprimitive matrix, 677
spheres, 275
splitting, 620
spring-mass vibrations, 570
springs, 86
square
  matrix, 8
  system, 5
  wave function, 301
stable, 544
  algorithm, 217, 317, 347, 422
  matrix, 544
  system, 544, 609
standard
  basis, 194, 240, 299
  coordinates, 240
  deviation, 296
  inner product, 95, 271
  scores, 296
standardization of data, 296
stationary distribution, 531, 693

steady-state distribution, 531, 636
steepest descent, 657
step size, 19
Stewart, G. W., xii
Stiefel, Eduard, 656
stiffness
  constant, 86
  matrix, 87
stochastic matrix, 685, 687
  doubly, 702
  summability of, 697
  unit eigenvalues of, 696
Strang, Gilbert, xii
strongly connected graph, 209, 671
Strutt, John W., 550
stuff in a vector space, 197, 200
subgroup, 402
submatrix, 7
  as a block matrix, 111
  and rank, 215
subscripts, 7
subset, 162
subspace, 162
  angles or gaps between, 450
  dimension of, 198
  directed distance between, 453
  four fundamental, 169
  invariant, 259, 262, 263
  maximal angle between, 455
  sum of, 205
substochastic matrix, 685
successive displacements, 623
successive overrelaxation method, 624
sum
  of matrices, 81
  of orthogonal projectors, 441
  of projectors, 393
  of vector spaces, 166, 383
    dimension of, 205
summable matrix and summability, 631, 633, 677
  stochastic matrices, 697
superdiagonal, 575
SVD, 412
  and full-rank factorization, 634
  and oblique projectors, 634
switching circuits, 539
Sylvester, James J., 44, 80, 411
Sylvester's law of inertia, 568
Sylvester's law of nullity, 220
symmetric
  functions, 494
  matrices, 85
    diagonalization and eigen components of, 549
    reduction to tridiagonal form, 352
    space of, 436
  permutation, 671

T

Taussky-Todd, Olga, 497
Taylor series, 18, 570, 600
t-digit arithmetic, 21
tensor product, 380, 597
  and the Laplacian, 573
term-by-document matrix, 419
text mining, 419
three-dimensional rotations, 328, 330
time domain, 363
Todd, John, 497
Toeplitz matrices, 514
Toeplitz, Otto, 514
total least squares, 223
trace, 90
  and characteristic equation, 504
  of imprimitive matrices, 678
  inequalities, 293
  of a linear operator, 256
  of a product, 110, 114
  of a projector, 392
  as sum of eigenvalues, 494
transformation, linear, 238
transient behavior, 532
transient class, 695
transition diagram, 108, 531
transition matrix, 108, 531, 688
transitive operations, 257
translation, in 3-D graphics, 332
transpose, 83
  and determinants, 463
  nullspace, 177
  properties of, 84
  reverse order law for, 109
  and similarity, 596
trapezoidal form, 342
trend of observations, 231
triangle inequality, 220, 273, 277
  backward version, 273
triangular matrices, 41, 103
  block versions, 112
  determinant of, 462
  eigenvalues of, 501
  elementary, 142
  inverses of, 122
triangularization, simultaneous, 522
triangularization using elementary reflectors, 342
triangular system, 6
tridiagonal matrix, 20, 156, 352
  Toeplitz matrices, 514
trivial
  nullspaces, 175
  solution, 57, 60, 69
    and nonhomogeneous systems, 70
    and nonsingular matrices, 116
  subspace, 162, 197
Tukey, J. W., 368, 375, 651

two-point boundary value problem, 18

U

unbiased estimator for variance, 449, 446
uncertainties in linear systems, 414
underrelaxation, 624
unique solution
  for differential equations, 541
  and free variables, 61
  for homogeneous systems, 61
  for nonhomogeneous systems, 70
unitarily invariant norm, 425, 337
unitary diagonalization, 547
unitary matrices, 304, 320
  determinant of, 473
unit columns, 102, 107
unit eigenvalues of stochastic matrices, 696
units, 27
unit sphere, 275
  image of, 414, 425
unstable, 544
upper-trapezoidal form, 342, 344
upper triangular, 103
URV factorization, 406, 407
  and full-rank factorization, 634

V

Vandermonde, Alexandre-Theophile, 185
Vandermonde determinant, 486
Vandermonde matrices, 185, 230, 357
Van Loan, Charlie, xii
variance, 447
vector, 159
  norms, 274
  spaces, 160
vertex matrix, 330
vibrations, small, 559
volume
  by determinants, 468
  by Gram–Schmidt, and QR, 431
von Mises, R., 533
von Neumann, John, 289

W

Weierstrass, Karl Theodor Wilhelm, 589, 662
well conditioned, 33, 127, 415
Weyl, Hermann, 160
why least squares?, 446
Wielandt, Helmut, 534, 666, 675, 679
Wielandt's matrix, 685
Wielandt's theorem, 675
Will, Marianne, xii
wire frame figure, 330
Woodbury, M., 124
Wronskian, 474, 481, 486
Wronski, Jozef M., 189
Wronski matrix, 189, 190

X, Y, Z

Young, David M., 625
Young, G., 411
Zeeman, E. Christopher, 704
zero nullspace, 175
zero transformation, 238
Z-matrix, 628, 639, 296
z-scores, 296